CUK-H05917-29-KP2899

THE JOURNAL OF PHILOSOPHY

29

VOLUME CVIII, NUMBER 1
JANUARY 2011

page
5 *What Is the Role of Consciousness in Demonstrative Thought?* Declan Smithies
35 *Does Probability Theory Refute Coherentism?* Michael Huemer

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027 Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2011 by the Journal of Philosophy, Inc ISSN 0022–362X

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA
SCIENCE CAMPUS

THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923. Since its founding, it has been published from the Department of Philosophy at Columbia University. It is typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 1
JANUARY–DECEMBER 2011

Edited by
David Albert, Bernard Berofsky, Akeel Bilgrami
John Collins, Arthur C. Danto, Jon Elster
Kent Greenawalt, Patricia Kitcher, Philip Kitcher
Charles Larmore, Isaac Levi, Wolfgang Mann
Frederick Neuhouser, Christopher Peacocke, Carol Rovane
Achille C. Varzi, Katja Vogt

Consulting Editors
James T. Higginbotham, Robert May
Charles D. Parsons, Wilfried Sieg

Managing Editor
Alyssa Timin

Published by The Journal of Philosophy, Inc.

Calcutta University Library
ALIPORE
CAMPUS

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple copying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing (e-mail preferred), and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all copying. Inquiries about copying may be sent to the Managing Editor. Please consult our website, www.journalofphilosophy.org, for more information about our copyright policies.

© Copyright 2011 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and other mailing offices. Please send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. Fax (212) 932-3721

KP 2899

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 1, JANUARY 2011

WHAT IS THE ROLE OF CONSCIOUSNESS IN DEMONSTRATIVE THOUGHT?*

Perception plays a fundamental role in enabling us to think about objects and properties in the world around us. In particular, perception of objects and properties enables us to think *demonstrative* thoughts about them.[1] But what must perception be like in order to play this role? Normal human perception is conscious in the sense that there is something it is like for us to perceive. This raises a question about the role of consciousness, which is the target question of this paper. Must perception be conscious if it is to enable us to think demonstrative thoughts about the world around us? And if so, why?

We can make the target question more vivid by considering examples of unconscious perception, such as blindsight.[2] Patients with blindsight are able to respond to stimuli that they do not consciously perceive. For example, when presented with a stimulus in the blind field and asked to guess whether it is an "X" or an "O," they answer correctly with a high degree of reliability. What explains this reliable ability is the patient's unconscious perception of the stimulus. So we can ask: is it

---

* Early versions of this paper were presented to the Basic Knowledge Workshop at the Australian National University in May 2007, the Annual Conference of the Australasian Association of Philosophy in July 2007, and the Workshop on Attention: Fundamental Questions at University College Dublin in May 2008. I am grateful to many people for helpful comments, including John Campbell, David Chalmers, Imogen Dickie, Chris Mole, James Pryor, Johannes Roessler, Susanna Siegel, Daniel Stoljar, Nico Silins, Wayne Wu, and an anonymous referee. Some of these ideas descend from my Ph.D dissertation at New York University, so I also record earlier debts to Ned Block, Paul Boghossian, Bill Brewer, Christopher Peacocke, Jim Pryor, and Crispin Wright.

[1] Not all demonstrative thought is based on perception. For example, we can think demonstrative thoughts on the basis of memory and testimony, as well as perception However, this paper is concerned only with demonstrative thought that is perceptually based, although this restriction is left implicit in the main text.

[2] For an overview of the empirical literature, see Lawrence Weiskrantz, *Consciousness Lost and Found: A Neuropsychological Exploration* (New York: Oxford, 1997)

0022-362X/11/0801/005–034

possible for a blindsighted patient to think demonstrative thoughts on the basis of unconscious perception? If so, then perception need not be conscious in order to provide a basis for demonstrative thought.

As a matter of empirical fact, it seems that blindsighted subjects are not able to think demonstrative thoughts on the basis of unconscious perception alone. For example, when presented with a stimulus in the blind field and asked what it is, they cannot answer. They respond to stimuli in the blind field only when they are forced to choose from a range of options. Even then, they claim to be only guessing and express surprise when informed of their reliability.[3] This suggests that the information provided by unconscious perception is not spontaneously accessible for use in thought, but only under the forced-choice conditions imposed by the experimenter. If this is right, then the subject must rely on more than unconscious perception in order to think about stimuli in the blind field. For instance, he may use the experimenter's prompting to infer that there are stimuli in his blind field on which he is being asked to report. In that case, he will be thinking about them by *description*, as the stimuli about which he is being asked.

However, these empirical observations do not settle the question at issue. I do not mean to be raising a purely empirical question about the role of conscious perception in demonstrative thought, but rather a distinctively philosophical question about whether there is a role for conscious perception that cannot be played by anything else. In other words, the question is not whether it is possible *in practice*, but whether it is possible *in principle* for blindsighted subjects to think demonstrative thoughts on the basis of unconscious perception. Therefore, we can abstract away from many of the contingent empirical facts about blindsight. In the spirit of Ned Block's "super-blindsighter," we might consider a counterfactual blindsighter whose unconscious perceptions are accessible for spontaneous use in the control of action, reasoning, and verbal report.[4] If the super-blindsighter is able to think demonstrative thoughts on the basis of unconscious perception, then we may conclude that perception need not be conscious in order to provide a basis for demonstrative thought.

The example of super-blindsight brings out another key point, which is that the relevant notion of consciousness is what Block calls *phenomenal consciousness*, according to which a state is conscious just in case there is something it is like for the subject to be in that state. This may be contrasted with Block's functionally defined notion of *access consciousness*, according to which a state is conscious just in case it is accessible for spontaneous use in the control of action, reasoning, and verbal report. The original point of Block's super-blindsight example was to illustrate the conceptual possibility of perception that is access conscious but not phenomenally conscious. The question that is prompted by the super-blindsight example is whether perception must be conscious in the phenomenal sense if it is to provide a basis for demonstrative thought.

---

[3] "When he was shown his results he [patient DB] expressed great surprise and insisted several times that he thought he was just 'guessing.'" See Weiskrantz, E. K. Warrington, M. D. Sanders, and J. Marshall, "Visual Capacity in the Hemianopic Field following a Restricted Occipital Ablation," *Brain*, xcvii (1974): 709–28, at p. 721

[4] Ned Block, "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences*, xviii, 2 (June 1995): 227–87.

My main aim in this paper is to argue for a positive answer to this question. First, though, I want to prepare the ground for my own account of the role of conscious perception in demonstrative thought by considering the alternative accounts proposed by Gareth Evans in *The Varieties of Reference* and John Campbell in *Reference and Consciousness*.[5] As we shall see, they agree with me that conscious perception is necessary for demonstrative thought, but they disagree with me about *why* it is necessary. Therefore, it will be instructive to compare and contrast these theories, since they offer importantly different accounts of the role of conscious perception in demonstrative thought. I begin with Evans, whose seminal work provides a natural starting point for theoretical reflection on the nature of demonstrative thought and the role of perception in demonstrative thought.

I

What is demonstrative thought? Rather than assuming any particular theoretical account from the outset, we can fix ideas with some examples. Suppose I am surrounded by white mugs and then blindfolded. In that case, I could point to one of the mugs at random and say, "That mug is white," but I would not thereby express a demonstrative thought about the mug. Rather, I would be thinking of it by description, as the mug to which I am now pointing. But if I were to remove the blindfold and take a look at the mug, then I would be in a position to think a demonstrative thought about the mug. As Evans puts the point, "Thinking about an object demonstratively is thinking about an object in a way which crucially depends upon the subject's currently *perceiving* that object."[6]

As this example illustrates, there is a distinction between descriptive and demonstrative thought about an object. Thinking demonstratively about an object is not just a matter of thinking about it by description, as the object that one is currently perceiving. The crucial difference is captured by Evans's observation that one cannot think about an object demonstratively unless one currently perceives it. By contrast, one can think about an object—say, Bismarck—as the object that one is currently perceiving even if one is not in fact perceiving it. Evans makes a closely related point in quoting G. E. Moore: "Can we say 'that thing' = 'the thing at which I am pointing' or 'the thing to which this finger points' or 'the nearest thing to which this finger points'? No because the prop[osition] is not understood unless the thing in question is *seen*" (305n).

---

[5] Gareth Evans, *The Varieties of Reference* (New York: Oxford, 1982); John Campbell, *Reference and Consciousness* (New York: Oxford, 2002).

[6] Evans, *op. cit.*, p 72. Subsequent references to this work are given in the text throughout section I.

There are further differences. For example, one cannot think about an object by description as the object that one is currently perceiving without using the concepts of perception and the first person. But one can think about an object demonstratively without using and, indeed, without even having these concepts. Moreover, Evans argues that, in order to acquire the concepts of perception and the first person, one already needs to have demonstrative concepts in place. In particular, grasping the concept of perceptual experience requires a capacity for self-ascribing perceptual experiences to oneself, which draws upon the very same concepts, including demonstrative concepts, that are used in forming beliefs about the external world (227).

Finally, if I perceive multiple objects, then there will be no unique object that satisfies the description "the object that I am now perceiving." To secure uniqueness, I will have to flesh out the description by means of further properties. But what if I am perceiving a collection of qualitatively identical but numerically distinct objects? In that case, I can distinguish them only by their perceived location. But in many cases, as in Evans's case of a bottle of pills spilled on a tabletop, our perceptual ability to distinguish places depends upon our perceptual ability to distinguish the objects located at those places. Thus, he writes, "The Idea of a point p in egocentric space, precise enough to be adequate to individuate the pill, exists only because there is something at p—the pill—for the subject's perception to latch on to...the Idea of p depends upon the perception of the pill, and hence is equivalent to the Idea 'where that pill is'" (172–73).

Demonstrative thought cannot be reduced to descriptive thought about the objects of one's demonstrations. As Evans puts it, "This gets things completely the wrong way round: it is the fact that I have my gaze fixed upon the thing, not the idea that I have my gaze fixed upon something, that determines which object is the object of my thought" (173). Here, Evans insists on the explanatory role of perception in enabling demonstrative thought: it is my perception of the object that explains how I am able to think demonstratively about that object. But how does perception play this explanatory role? Evans's proposal is that perception enables us to think demonstrative thoughts about objects by providing us with information about the object which is causally derived from the object in question. When someone thinks a demonstrative thought about an object, he exploits an information link with the object. In this sense, demonstrative thoughts are *information-based* thoughts.[7]

Evans argues that the mere existence of a perceptual information link with an object is not sufficient to enable demonstrative thought about the object in the absence of certain psychological capacities, including the capacity to form beliefs about the object, to locate the object in space, and to keep track of its movements through space. Arguably, the second and third conditions are too demanding, since one can think demonstrative thoughts about an object whose location is misperceived or whose movements are too fast and sudden for perceptual tracking. However, the first condition seems necessary if one's information link is to enable demonstrative thought about the object in question. To flesh out the proposal in more detail, one has a demonstrative concept or way of thinking of an object only if one has perceptual information from the object which one is disposed to use in forming beliefs about the object, without mediation by any background beliefs. As Evans formulates this condition, "[A] subject who has a demonstrative Idea of an object has an *unmediated* disposition to treat information from that object as germane to the truth and falsity of thoughts involving that Idea" (146).

If this is the role that perceptual information plays in enabling demonstrative thought, we can ask: what must perceptual information be like in order to play this role? Evans claims that perceptual information is belief independent, in the sense that it is independent of the subject's beliefs about the world, and that it is nonconceptual, in the sense that it does not involve the exercise of conceptual abilities of the kind that are exercised in belief formation. However, he nowhere claims that perceptual information must be conscious. Indeed, he explicitly denies this:

> So far I have been considering the non-conceptual content of perceptual informational states. Such states are not *ipso facto* perceptual *experiences*—that is, states of a conscious subject. However addicted we may be to

---

[7] In the category of information-based thought, Evans includes not only demonstrative thought, but also recognitional thought and thought involving proper names. The distinctive feature of demonstrative thought is that it requires a current information link with the object of thought.

> thinking of the links between auditory input and behavioural output in information-processing terms—in terms of computing the solution to simultaneous equations—it seems abundantly clear that evolution could throw up an organism in which such advantageous links were established, long before it had provided us with a conscious subject of experience. (157–58)

Evans mentions blindsight as an example of unconscious perceptual information. The question, then, is whether there is anything in Evans's theory to exclude the possibility that a blindsighted subject could form demonstrative thoughts on the basis of unconscious perceptual information. His theory appeals to the existence of an information link and its functional role in enabling various psychological capacities, such as locating, tracking, and forming beliefs about an object. On the face of it, however, there is nothing in Evans's theory to rule out the possibility that unconscious perceptual information plays the requisite functional role. In fact, the situation is rather more complicated. Evans writes:

> [W]e arrive at conscious perceptual experience when sensory input is not only connected to behavioural dispositions in the way I have been describing—perhaps in some phylogenetically more ancient part of the brain—but also serves as the input to a *thinking, concept-applying, and reasoning system*; so that the subject's thoughts, plans and deliberations are also systematically dependent on the informational properties of the input. When there is such a further link, we can say that the person, rather than just some part of his brain, receives and possesses the information. (158)

Here, Evans sketches a broadly functionalist theory of consciousness.[8] The claim is that perceptual information is conscious if it serves as input to a thinking, concept-applying, and reasoning system—that is, by the formation of demonstrative and other information-based thoughts. In other words, it is a sufficient condition for perceptual information to be conscious that it serves as an input for demonstrative thought. Equivalently, it is a necessary condition for perceptual information to serve as an input for demonstrative thought that it is conscious. Hence, Evans's functionalist theory of consciousness rules out the possibility that a blindsighter could form demonstrative thoughts on the basis of unconscious perceptual information. A blindsighter who thinks demonstrative thoughts would be a kind of partial functional zombie, since his unconscious perceptions would play the functional role of our own conscious perceptions. However, the possibility of functional zombies is ruled out by any functionalist theory of consciousness.

---

[8] Compare Michael Tye's PANIC theory, which identifies phenomenal character with Poised Abstract Nonconceptual Intentional Content, that is, nonconceptual representation which is poised to make an impact on the conceptual system. See Tye, *Ten Problems of Consciousness* (Cambridge: MIT, 1995).

How plausible is Evans's functionalist theory of consciousness? A functionalist theory of consciousness aims to explicate the nature of consciousness by specifying functional conditions that are both necessary and sufficient for consciousness. However, Evans's theory is incomplete, since it provides only sufficient conditions. A standard objection to sufficiency is that functional zombies are possible, but this objection lacks dialectical force in the present context, since what is at issue is precisely whether there could be a certain kind of partial functional zombie—namely, a blindsighted subject who thinks demonstrative thoughts on the basis of unconscious perceptual information. Nevertheless, we can ask whether Evans's theory explains *why* a functional zombie is impossible by considering how his functionalist theory of consciousness might be elaborated to yield conditions that are both necessary and sufficient.

The obvious proposal is that what makes perceptual information conscious is the fact that it serves as input to the conceptual system for use in demonstrative thought. However, the equally obvious objection is that this is an over-intellectualization, since conceptual capacities are not necessary for consciousness. Evans himself objects to higher-order thought theories of consciousness on the grounds that "experience can antedate thoughts about it," but he continues, "although it is true that our intuitive concept requires a subject of experience to have *thoughts*, it is not thoughts about the experience that matter, but thoughts about the world" (158). Yet, surely experience antedates not only thoughts about experience but also thoughts about the world? Indeed, Evans himself imposes the *generality constraint*, according to which the conceptual capacities involved in thought must be systematically recombinable (100–05). However, it is unlikely that cognition in higher animals, such as chimpanzees, is sufficiently systematic to meet these conditions for conceptual thought.[9] Presumably, though, it would be grossly implausible to deny that these animals are subjects of conscious perceptual experience.

An alternative proposal is that what makes perceptual information conscious is the fact that it serves as input to a central cognitive system, which may be either a conceptual or a nonconceptual system. On this proposal, the experiences of nonhuman animals need not

[9] For several compelling examples, see Susan Hurley, "Making Sense of Animals," in Hurley and Matthew Nudds, eds., *Rational Animals?* (New York: Oxford, 2006), pp. 139–71

serve as input to a conceptual system for use in thought, so long as they serve as input to a nonconceptual cognitive system for use in other executive functions, such as the control of action. This proposal avoids the objection to necessity but only at the cost of inviting an objection to sufficiency. For while it may be sufficient for perceptual information to be conscious that it serves as input for demonstrative thought, it is not sufficient that it serves as input to a central cognitive system that plays a role in the control of action. On the face of it, there is nothing to rule out the possibility of a central cognitive system that is wholly comprised of unconscious informational states.[10]

A further objection is that these functionalist proposals threaten to undermine the causal-explanatory role of consciousness.[11] Conscious perceptual experience does seem to play a causal-explanatory role in enabling us to think demonstrative thoughts about the world around us, although it is a further question whether its role is indispensable. Assuming functionalism, however, the fact that one's perceptual information is conscious plays no substantial role in the causal explanation of demonstrative thought. Rather, the fact that it is conscious is merely a consequence of its role in the explanation of demonstrative thought, since what it is for perceptual information to be conscious is just for it to serve as an input for demonstrative thought. Therefore, consciousness is reduced to the status of a mere epiphenomenon in the causal explanation of demonstrative thought.

However, one might accept Evans's claim that there is a necessary connection between consciousness and demonstrative thought even if one rejects his explanation in terms of a functionalist theory of consciousness. The proposed explanation is that it is essential to the nature of *consciousness* that perceptual information is conscious if it serves as an input for demonstrative thought. But there is an alternative explanation, according to which it is essential to the nature of *demonstrative thought* that it is formed on the basis of perceptual information that is conscious. In other words, the necessary connection between consciousness and demonstrative thought may derive not from the nature of consciousness, but rather from the nature of demonstrative thought.[12]

---

[10] Consider the role of cognitive maps in insect navigation, which integrate perceptual information with stored information for use in the control of action. For an overview, see Charles Gallistel, *The Organization of Learning* (Cambridge: MIT, 1990).

[11] See also Campbell, *op. cit.*, chapter 7.

[12] Kit Fine argues that claims of essence, unlike claims of necessity, are sensitive to source. See Fine, "Essence and Modality," in James Tomberlin, ed., *Philosophical Perspectives 8 Logic and Language* (Atascadero, CA: Ridgeview, 1994), pp. 1–16.

II

In *Reference and Consciousness*, Campbell argues that conscious perceptual attention to an object is necessary for thinking demonstrative thoughts about the object. He argues that grasping a demonstrative concept of an object is a matter of knowing its reference—that is, knowing to which object it refers. And he argues that conscious perceptual attention to an object is what provides us with our knowledge of the reference of demonstrative concepts. So, like Evans, but for different reasons, he argues that blindsighted subjects cannot think demonstrative thoughts on the basis of unconscious perceptual information. They cannot grasp demonstrative concepts of the objects in their blind field, since they do not know which objects are in question.

To make this point intuitively compelling, Campbell asks us to imagine what it is like to experience a "sea of faces":

> [Y]ou and I are sitting at a dinner table with a large number of people around and you make a remark to me about 'that woman'. There are a lot of people around; I can't yet visually single out which one you mean. So on anyone's account, I do not yet know which woman you are talking about. Suppose now that we add to the example. My visual experience remains as before: a sea of faces. I cannot consciously single out the person you mean. All I get consciously is the sea of faces. But now we add some of what the blindseer has.... So I can make reliable guesses about what the person is eating, wearing, and so on, as well as reaching and pointing appropriately. But so long as my conscious experience remains a sea of faces, there is an ordinary sense in which I do not know who you mean.[13]

The suggestion here is that it does not matter how much information you receive from the object or how you are able to use it in the control of action, reasoning, and verbal report. In the absence of conscious perceptual attention to the object, there is an intuitive sense in which you do not know which object is in question. But if you do not know which object is in question, then you cannot grasp demonstrative concepts of the object.

It may be objected that there is an intuitive sense in which you do know which object is in question.[14] After all, if you are in the same position as the super-blindsighter, then the perceptual information that you receive from the object enables you to locate the object and to track its movements through space, as well as to give reliable reports about its properties. Of course, the obvious response is that there remains an intuitive sense in which you do not know which object is in question, which is the sense that is relevant for grasping demonstrative concepts. However, it is not clear how much weight this response can bear in the absence of further theoretical considerations.

---

[13] Campbell, *op. cit.*, pp. 8–9. Subsequent references to this work are given in the text throughout section II.

[14] This objection is made by Sean Kelly, "Reference and Attention: A Difficult Connection," *Philosophical Studies*, CXX, 1–3 (2004): 277–86.

All sides can agree that the way in which the super-blindsighter thinks about the objects in his blind field is different from the way that you and I think about the objects that we experience. The question is what this difference amounts to. Campbell argues that our thoughts are genuinely demonstrative thoughts, whereas those of the super-blindsighter are merely descriptive thoughts. But the alternative is to say that we think demonstrative thoughts on the basis of conscious perceptual experience, whereas the super-blindsighter thinks demonstrative thoughts on the basis of unconscious perceptual information. The question is whether consciousness is an essential feature of demonstrative thought or whether it is merely an accidental feature of the way in which we happen to think these thoughts. To decide this question, we need to supplement the appeal to commonsense intuitions by means of a more fully developed theory of demonstrative thought.

Campbell himself would be the first to concede this point. After all, the main aim of *Reference and Consciousness* is to develop a theory of demonstrative thought which purports to explain why the role of consciousness is essential. The centerpiece is Campbell's account of the relationship between knowing the reference of a demonstrative concept and one's way of using the concept in thought. According to what he calls *the classical view*, knowing the reference of a demonstrative concept is what causes and justifies one's ways of using the concept in thought. The classical view provides a functional characterization of the state that constitutes knowing the reference of a demonstrative concept. This state, whatever it is, causes and justifies one's use of particular ways of verifying and acting upon propositions involving the demonstrative concept. Campbell goes on to argue that conscious perceptual attention is the state that occupies this functional role.[15] Therefore, he concludes that conscious perceptual attention to an object is what constitutes our knowledge of the reference of demonstrative concepts of the object. This line of argument is summarized in the following passage:

> Knowledge of the reference of the demonstrative is what causes and justifies the use of particular procedures to verify and find the implications of

---

[15] Campbell claims that conscious perceptual attention occupies this role only when certain background conceptual capacities are in place. Therefore, conscious perceptual attention does not constitute knowledge of the reference of demonstrative concepts in conceptually unsophisticated creatures.

> propositions containing the demonstrative. Conscious attention to the object, I will argue, is what causes and justifies the use of particular procedures for verifying and finding the implications of propositions containing the demonstrative. Hence, knowledge of the reference of the demonstrative is provided by conscious attention to the object. (25–26)

How exactly does conscious perceptual attention play its role in causing and justifying our use of demonstrative concepts in thought? Campbell locates its functional role at the computational level of information-processing mechanisms. He claims that conscious perceptual attention plays a role in setting in motion and defining the targets for the unconscious information processing that underpins our ways of using demonstrative concepts in thought. Suppose you want to perform an action or to verify a proposition about some demonstratively identified object. If you are to succeed in this task, then your information-processing mechanisms must solve certain computational problems. But in order to solve these computational problems, they must be able to select the right kind of unconscious information to be processed. How is this achieved? Campbell's proposal is that conscious perceptual attention plays a crucial role in the selection of information by identifying the object that is the target of the underlying information processing.[16]

Without going into all the details of Campbell's complex and sophisticated account, it may be helpful to consider an example. Suppose your task is to determine whether an object *x* is enclosed within the boundaries of a looped figure. How do we solve this task? According to Shimon Ullman, the visual system makes use of a kind of "coloring" operation, which spreads outwards from the target object *x* in all directions until it reaches a boundary; this operation yields the answer that the *x* is enclosed if and only if the coloring operation comes to an end.[17] Campbell's claim is that the role of conscious perceptual attention is to set the targets for this unconscious computational process by selecting the object *x* as the starting point for the coloring operation.

The role of conscious perceptual attention, then, is to define the targets for the unconscious information processing that underpins our use of demonstrative concepts in thought. But why can't this target-setting role be played by anything else? For example, in the super-blindsight example, there must be something remote from consciousness that plays the role of defining the targets for the information processing that underpins actions and verbal reports directed towards a particular object. Otherwise, we cannot explain the reliability of those actions and verbal reports. But then why suppose that it is essential for demonstrative thought that this target-setting role should be played by conscious perceptual attention, rather than by something entirely remote from consciousness?

---

[16] Compare David Milner and Melvyn Goodale's proposal that conscious visual processing in the ventral stream sets the targets for unconscious visual processing in the dorsal stream. See Milner and Goodale, "Attention, Consciousness, and the Coordination of Behaviour," in *The Visual Brain in Action* (New York: Oxford, 1995).

[17] Shimon Ullman, *High Level Vision: Object Recognition and Visual Cognition* (Cambridge: MIT, 1996).

Campbell's answer is that conscious experience of objects is necessary for thinking of objects in categorical terms, rather than merely in dispositional terms. Intuitively, we do not think of objects merely in terms of their dispositions to enter into certain sorts of causal relations. Rather, we think of objects as the categorical basis of their dispositions. Campbell claims that we cannot explain how we are able to think of objects in categorical terms except by appealing to the fact that conscious experience relates us to categorical objects themselves and not merely their associated functional dispositions. Moreover, this is what he takes to explain why subjects with blindsight are unable to think demonstrative thoughts about objects in their blind field. They do not represent the objects in their blind field as categorical, but merely in terms of their possession of certain dispositions. For instance, they represent the Gibsonian "affordances" that facilitate certain ways of acting on those objects, but they do not represent the categorical grounds of their affordances. This, according to Campbell, is precisely what is missing in blindsight:

> But what is the [blindsighted] subject missing? If the subject reaches and grasps successfully, the subject nonetheless does not know why the reaching and grasping has been successful. That is the real reason why such a subject, no matter how fast, accurate, and reliable she may become, is still said to be 'only guessing'. All the subject has is the fact of success in action, without any understanding of why she has been successful. The subject has been right in thinking that there are these affordances there, but does not know why the world has afforded just this and that. (144)

I will not dispute Campbell's claim that a blindsighted subject cannot think of the objects in his blind field in categorical terms, rather than merely in dispositional terms. However, it is not clear to me that this claim represents much of an advance on the initial intuition that the blindsighter cannot think about those objects demonstratively, but only by description—say, as the possessors of certain dispositional characteristics. The question at issue here is not whether this intuition is correct, but whether we can make any theoretical sense of it. My aim in the remainder of this section is to raise three challenges to the theoretical account that Campbell proposes.

Why is conscious experience necessary for thinking of objects in categorical terms? Campbell claims that we cannot explain our categorical conception of objects except by appeal to a *relational* view on which experience is a direct relation between the subject of experience and the categorical objects and properties in his environment. But why not appeal instead to a *representational* view on which experience is a relation between the subject of experience and a representational content? Campbell argues on the one hand that a representational view cannot explain why experience represents categorical objects and properties, as opposed to associated complexes of dispositions; on the other hand, he argues that this is exactly what a representational view of experience needs to explain. However, Campbell's argument overlooks the distinction between conceptual and nonconceptual forms of representation. If we assume that experience nonconceptually represents objects as categorical, then we can appeal to this fact in explaining how we conceptually represent objects as categorical in thought.[18] The first challenge, then, is to explain why we need a relational view of experience, as opposed to a representational view, in order to explain our categorical conception of objects.

The second challenge is to explain why we need a relational view of *experience*, as opposed to a relational view of *perception*, in order to explain our categorical conception of objects. Campbell argues that conscious experience explains our conception of objects as categorical by putting us in direct relation to categorical objects themselves. But why does this perceptual relation have to be conscious? On some empirical theories of vision, there are unconscious visual mechanisms that put us in direct relation to objects. Zenon Pylyshyn, for example, argues that early visual processing involves a limited number of visual indexes, or "FINSTs," which function by demonstrating objects and tracking their movements without representing their locations or any other properties.[19] This contrasts with a more orthodox picture, on which the representation of objects is constructed from more basic representations of properties by means of a process that binds properties together by their represented locations.[20] It is, of course, an empirical question which of these views best describes the actual workings of the visual system. However, if Pylyshyn's view is even possibly true, then it seems to provide a counterexample to the claim that conscious experience is necessary for demonstrative, as opposed to descriptive, reference to objects. The second challenge, then, is to explain why conscious perceptual *experience* is necessary for demonstrative reference.

---

[18] For a version of this proposal, see Christopher Peacocke, "Does Perception Have a Nonconceptual Content?" this JOURNAL, XCVIII, 5 (May 2001): 239–64.

[19] Zenon Pylyshyn, "Visual Indexes, Preconceptual Objects and Situated Vision," *Cognition*, LXXX, 1 (June 2001): 127–58.

[20] Anne Treisman and Gary Gelade, "A Feature Integration Theory of Attention," *Cognitive Psychology*, XII, 1 (January 1980): 97–136.

The third and final challenge is to explain why conscious perceptual *attention* is necessary for demonstrative thought. Campbell uses the Sea of Faces example to prompt the intuition that conscious experience of an object is not sufficient for demonstrative thought about the object unless one is consciously attending to the object. But the question is whether his theoretical account of demonstrative thought succeeds in explaining this intuition. Campbell imposes two conditions that perception must meet to provide a basis for demonstrative thought. First, perception must be conscious to explain our conception of objects as categorical by putting us in relation with categorical objects themselves, rather than merely their associated dispositions. Second, perception must involve attention in order to play its functional role in defining the targets for the information processing that underpins our use of demonstrative concepts in thought. Campbell's proposal is that only conscious perceptual attention satisfies both of these conditions at once.

The problem is that both of Campbell's conditions are satisfied in the Sea of Faces case, but in the absence of conscious perceptual attention. First, the subject's conscious experience relates him to various objects in the scene, which nevertheless do not engage his conscious attention. Second, his conscious experience plays the right kind of functional role in defining the targets for the information processing that underpins his actions and verbal reports concerning the objects in question. Intuitively, however, the subject is unable to think demonstrative thoughts about these objects, so long as his experience remains a sea of faces. Therefore, Campbell's proposal fails to explain the intuitions that the Sea of Faces example was designed to elicit.

The Sea of Faces case illustrates an important contrast between two concepts of attention which should be carefully distinguished.[21] In much recent empirical work, attention is operationally defined in terms of its functional role in selecting information for certain kinds of processing. In this functional sense, subjects with blindsight may attend to objects in the blind field.[22] However, the functional concept of attention as selection of information from an object may be contrasted with the phenomenal concept of attention as a distinctive way of being conscious of an object. In this sense of attention, there is a phenomenal contrast between objects in the attended foreground and objects in the unattended background of conscious experience. Campbell uses the metaphor of "experiential highlighting" to convey this phenomenal concept of attention:

---

[21] For further discussion of the distinction between phenomenal and functional concepts of attention, see Declan Smithies, "Attention as Rational-Access Consciousness," in Christopher Mole, Smithies, and Wayne Wu, eds., *Attention: Philosophical and Psychological Essays* (New York: Oxford, 2011), pp 247–73.

[22] See Robert Kentridge, Charles Heywood, and Weiskrantz, "Attention without Awareness in Blindsight," *Proceedings of the Royal Society*, B266 (1999): 1805–11.

> If I am to understand a demonstrative referring to an object, it is not enough merely that the object be there somewhere in my visual field; I have to attend to it. But the attention that is needed here is, as it were, a matter of experiential highlighting of the object; it is not enough that there be some shifts in the architecture of my information-processing machinery, remote from consciousness. (2)

In the Sea of Faces case, the functional and phenomenal concepts of attention come apart. There is functional selection of information from the target object, but there is no experiential highlighting of the target object. Intuitively, what is necessary for demonstrative thought about an object is not mere selection of information from the object, but experiential highlighting. To explain this, however, we need to look elsewhere.

III

My aim in what follows is to propose an alternative to Campbell's account of the role that consciousness plays in explaining our grasp of demonstrative concepts. Following Campbell, I will assume that consciousness explains our grasp of demonstrative concepts by virtue of its role in causing and justifying our ways of using those concepts in thought. However, I will offer a distinctive account of why the role of consciousness is essential. On Campbell's account, consciousness plays an essential *target-setting* role: it sets in motion and defines the targets for the information processing that underpins our use of demonstrative concepts in thought. On my account, by contrast, consciousness plays an essential *epistemic* role: it enables subjects to use demonstrative concepts in forming immediately justified beliefs about objects in the world around them.

The crucial difference between these two accounts is that they are located at different levels of psychological explanation. Campbell's account is located at what we might call the *computational* level, whereas mine is located at the *rational* level.[23] Rational and computational processes are in many ways alike. Both are intentional processes, which involve causal transitions between intentional states that are causally sensitive to the intentional contents of those states. However, rational and computational processes are subject to different kinds of normative evaluation. Rational processes, including deductive and inductive inference, as well as noninferential modes of belief revision, are subject to evaluation in terms of distinctively epistemic norms of rational justification. Computational processes, on the other hand, are subject to evaluation in terms of distinctively computational norms of proper functioning. It does not make sense to ask whether one's visual system is rationally justified in performing a computation that yields a visual representation of the distal environment or fine-tunes the spatial parameters of a visually guided action.[34]

---

[23] Compare Daniel Dennett's related distinction between personal and subpersonal levels of explanation. See Dennett, *Content and Consciousness* (London: Routledge, 1969).

It is consistent with my proposal that consciousness plays an important functional role at the computational level as well as the rational level. There is certainly no commitment to the claim that these levels are causally isolated from one another. It is more likely that consciousness serves as a common cause of psychological processes at both levels in such a way that its epistemic role at the rational level is underpinned by its target-setting role at the computational level. Nevertheless, we should distinguish between two questions: (1) What role is actually played by conscious perceptual attention? (2) What role could only be played by conscious perceptual attention? In the previous section, I argued that the target-setting role of consciousness could have been played by something entirely remote from consciousness; in this section, I will argue that consciousness is essential for playing an epistemic role, which could not be played by anything else.

My overall strategy will be to argue from two premises: the first premise articulates an epistemic condition for grasping demonstrative concepts, while the second premise makes a claim about the epistemic role of consciousness. Together, these two premises yield the conclusion that consciousness plays an essential role in explaining our grasp of demonstrative concepts. In this section, I will develop the argument for the role of conscious perceptual experience; in the next section, I will consider whether the argument can be generalized in such a way as to extend to the role of conscious perceptual attention. The basic argument can be summarized as follows:

---

[34] Note that Campbell uses the term "justification" in a broad sense that includes both rational and computational norms, but this should not obscure the distinction between these two different subspecies of norms.

The Experience Argument

(1) One has a demonstrative concept of an object *o* only if one has information about *o* which provides immediate, defeasible justification to form beliefs about *o*.
(2) One has information about *o* which provides immediate, defeasible justification to form beliefs about *o* only if one has conscious perceptual experience of *o*.
(3) One has a demonstrative concept of an object *o* only if one has conscious perceptual experience of *o*.

The first premise is an instance of a more general thesis about the epistemic constraints on concept possession. What is it to possess a concept? To have a concept is to have a kind of mental representation.[25] But not just any kind of mental representation counts as a concept. For example, David Marr's computational theory of vision posits mental representations in the visual system which represent such things as edges, blobs, and zero-crossings.[26] And yet there is no reason to suppose that anyone has the concept of a zero-crossing unless they are already familiar with Marr's work. So, what is it to have the concept of a zero-crossing? An obvious answer is that to have the concept of a zero-crossing is to have a capacity for thinking about zero-crossings as such.[27] But this merely defers the question: what is it to think about zero-crossings as such if it is not enough to have mental representations in early vision which represent zero-crossings?

On the assumption that mental representations are functionally individuated, we can distinguish concepts from other mental representations by their distinctive functional role in our mental lives. But what is the distinctive functional role of a concept? A plausible answer is that concepts are mental representations that are epistemically individuated by their disposition to play a role in rational processes. Mental representations in early vision are not concepts because they are disposed to play a role in computational processes, which are not subject to evaluation in terms of the distinctively epistemic norms that apply to rational processes.[28]

---

[25] Concepts are sometimes construed not as mental representations, but as abstract objects to which we are related by our conceptual mental representations. But this is merely a terminological issue, since what I have called "concepts" might equally be called "conceptual mental representations."

[26] David Marr, *Vision* (New York: Freeman, 1982).

[27] Jerry Fodor, *Concepts: Where Cognitive Science Went Wrong* (New York: Oxford, 1998).

[28] For this reason, my claim that consciousness is essential for our grasp of demonstrative concepts is compatible with Pylyshyn's claim that there are demonstrative mental representations in early vision.

KP2899

We can formulate the epistemic constraints on concept possession in terms of a *possession condition*, which specifies the dispositions that a subject must have as a condition for possessing the concept.[29] If concepts are individuated in epistemic terms, then the dispositions mentioned in the possession condition for a concept must be rational dispositions. However, it does not follow that all of our dispositions to use a concept are rational dispositions, since not all of our dispositions to use a concept need enter into the individuation of the concept. In particular, rational dispositions are privileged over irrational dispositions in such a way that only rational dispositions enter into the individuation of a concept. This explains why we can make sense of sharing our concepts with a perfectly rational thinker, who has no irrational dispositions, although we cannot so easily make sense of sharing our concepts with a perfectly irrational thinker, who has no rational dispositions.[30]

What are the possession conditions for demonstrative concepts? According to Evans, one has a demonstrative concept of an object only if one has perceptual information from the object which one is disposed to use in forming beliefs about the object without mediation by any background beliefs. If this is a rational disposition, we may conclude that one's perceptual information from the object provides one with immediate justification to form beliefs about the object. Otherwise, it would be irrational for one to form beliefs on the basis of perceptual information without mediation by background beliefs. Thus, one has a demonstrative concept of an object only if one has perceptual information from the object, which provides one with immediate, defeasible justification to form beliefs about the object. This is enough to establish our first premise.

So far, I have been arguing that one's perception of an object must play an epistemic role in order to provide one with a demonstrative concept of the object. Now we can ask: what must one's perception be like in order to play this epistemic role? Here, I claim that the phenomenology of conscious perceptual experience plays an essential epistemic role. It is because my experience has the phenomenology of confronting me with objects and properties in the world around me that it causes me to form beliefs about the world around me and justifies me in doing so. Moreover, my justification has its source in the phenomenology of experience, rather than in background beliefs about the reliability of my experience. After all, it is not clear what could justify my background beliefs about the reliability of my experience besides my experiences themselves. Of course, the justification provided by my experience is *defeasible* in the sense that it can be defeated by justification to believe that my experience is unreliable. But in the absence of defeaters, my justification to take experience at face value is *immediate* in the sense that it does not depend on my having independent justification for background beliefs about the reliability of my experience.[31]

---

[29] See Peacocke, *A Study of Concepts* (Cambridge: MIT, 1992), chapter 1.

[30] Compare Ralph Wedgwood, "The Normativity of the Intentional," in Brian McLaughlin and Ansgar Beckermann, eds., *The Oxford Handbook of the Philosophy of Mind* (New York: Oxford, 2009), pp. 421–36, and "Normativism Defended," in McLaughlin and Jonathan Cohen, eds., *Contemporary Debates in the Philosophy of Mind* (Malden, MA: Blackwell, 2007), pp. 85–101.

The epistemic role of consciousness is best illustrated by reflection on cases in which consciousness is missing. Consider the epistemic predicament of subjects with blindsight. Do subjects with blindsight have immediate justification to form beliefs on the basis of unconscious perceptual information about objects in the blind field? What goes missing in blindsight is any phenomenology of confrontation with objects and properties in the blind field. This is why subjects with blindsight do not spontaneously form beliefs about objects in the blind field. Under forced-choice conditions, they are reliable in certain kinds of discrimination tasks, but they tend to regard their responses as mere guesswork and express surprise when informed of their reliability. From their own subjective perspective, there is nothing to distinguish their verbal reports about the blind field from mere guesswork. Intuitively, then, it is no more rational for them to form beliefs about the blind field than it is to form beliefs on the basis of blind guesswork.

Reflection on blindsight suggests that the mere reliability of one's information link with an object is not sufficient to provide one with justification to form beliefs about the object. Compare Laurence BonJour's well-known example of the clairvoyant, Norman, who forms beliefs about the location of the President on the basis of a reliable clairvoyant faculty, although he lacks any justification to believe that he has this reliable faculty.[32] Intuitively, despite their reliability, Norman's beliefs are no more justified than beliefs formed on the basis of blind guesswork. This is because there is nothing accessible from Norman's own subjective perspective to distinguish his clairvoyant beliefs from beliefs formed on the basis of guesswork. And crucially, as BonJour notes, "the rationality or justifiability of Norman's belief should be judged from Norman's own perspective rather than from one which is unavailable to him."[33]

---

[31] This is what James Pryor has called a "dogmatist" theory of justification. Crispin Wright has argued for an opposing view, on which justification from experience depends upon independent justification to believe (or, in his terms, entitlement to accept) that experience is reliable, which we have *by default*. This view is also consistent with an epistemic asymmetry between conscious and unconscious perception such that we have default justification to believe that conscious perception is reliable, but we need to earn justification to believe that unconscious perception is reliable. See Pryor, "The Skeptic and the Dogmatist," *Noûs*, xxxiv, 4 (December 2000): 517–49; and Wright, "Warrant for Nothing (and Foundations for Free)?" *Supplement to the Proceedings of the Aristotelian Society*, lxxviii, 1 (July 2004): 167–212.

[32] Laurence BonJour, *The Structure of Empirical Knowledge* (Cambridge: Harvard, 1985), p. 41.

The argument here rests on a key epistemological assumption, which is that justification depends only on facts that are accessible from one's own subjective perspective or point of view on the world. But why should we accept this assumption? A plausible answer is that these are the only facts that one has to go on in critical reflection about what to believe. So, the underlying rationale for this assumption is that the concept of justification (and, hence, the property that it denotes) is essentially tied to its role in regulating the practice of critical reflection.[34] Roughly speaking, we can say that a belief is justified if and only if it has what it takes to survive critical reflection. Therefore, we can argue that Norman's clairvoyant belief is unjustified because it does not have what it takes to survive critical reflection. Compare BonJour's own diagnosis of the case:

> Norman's acceptance of the belief about the President's whereabouts is epistemically irrational and irresponsible, and thereby unjustified, whether or not he believes himself to have clairvoyant power, so long as he has no justification for such a belief. Part of one's epistemic duty is to reflect critically upon one's beliefs, and such critical reflection precludes believing things to which one has, to one's knowledge, no reliable means of epistemic access.[35]

Blindsighted subjects are in the same epistemic predicament as the clairvoyant, Norman. They have a reliable perceptual faculty which enables them to make verbal reports about objects in the blind field on the basis of unconscious perceptual information. However, since this perceptual information is unconscious, they do not have justification to believe in their own reliability, which is why they regard their responses in forced-choice tasks as mere guesswork. Of course, these subjects eventually learn that they are reliable through induction and testimony, in which case they do have justification to form beliefs about objects in the blind field, but only because of their background justification to believe in their own reliability. So, there is a dilemma in the offing: either they are justified by background beliefs about their own reliability, or they are not justified at all. Therefore, unconscious perception is not sufficient to provide a source of immediate justification to form beliefs about objects in the blind field.

---

[33] *Ibid.*, pp. 43–44.

[34] This conception of justification is motivated and developed in much more detail in forthcoming work. See Smithies, "Why Care about Justification?" (manuscript) and "Moore's Paradox and the Accessibility of Justification," *Philosophy and Phenomenological Research* (forthcoming).

[35] BonJour, *op. cit.*, p. 42.

In summary, there is an epistemic asymmetry between conscious perceptual experience and unconscious perception in blindsight. Moreover, we cannot explain this asymmetry except in terms of the presence or absence of consciousness. In other words, perception of an object provides immediate justification to form beliefs about the object only if it is conscious. It may be objected that blindsight, unlike ordinary perception, involves the feeling of guessing, which is a defeater for any immediate perceptual justification that is provided in blindsight. But this fails to account for the super-blindsighter, who forms beliefs spontaneously and confidently on the basis of unconscious perceptual information about objects in the blind field.[36] In this case, there is no feeling of guessing, and yet there is nothing besides the feeling of confidence to distinguish the beliefs in question from beliefs formed on the basis of guesswork. However, the mere feeling of confidence is not sufficient to justify anyone's belief—justification is not so easy to come by!

A different suggestion is that what is missing in blindsight is second-order knowledge of the causal basis of one's beliefs.[37] In other words, ordinary perceivers know when they form beliefs on the basis of perceptual experience, whereas subjects with blindsight or super-blindsight cannot know the causal basis of their beliefs. However, this fails to account for the epistemic asymmetry between unconscious perception in blindsight and conscious perceptual experience in unreflective subjects. Evidence from developmental psychology suggests that three-year-old children do not understand whether their beliefs are formed on the basis of perception, inference, or testimony.[38] However, it is surely an over-intellectualization to deny that their beliefs can be immediately justified on the basis of perceptual experience. Arguably, these children are in an epistemic position to know the causal basis of their beliefs, although they lack the psychological capacities to take advantage of this epistemic position. But if so, they are in an epistemic position to know the causal basis of their beliefs only because their beliefs are formed on the basis of conscious perceptual experience.

---

[36] Block (*op. cit.*, p. 285) states that the super-blindsighter is "trained to prompt himself at will, guessing without being told to guess," but suppose instead that he forms beliefs spontaneously on the basis of unconscious perception, without any need for prompting himself to guess. This undermines Tye's claim (*op. cit.*, pp. 142–43) that visual representations in super-blindsight are not poised in the right way for PANIC content because they impact on the belief system only indirectly via an act of willing.

[37] Michael Ayers suggests that blindsight involves knowing without knowing how one knows: in his terms, this is secondary knowledge, but not primary knowledge. See Ayers, *Locke: Epistemology and Ontology* (New York: Routledge, 1991).

[38] Alison Gopnik and Peter Graf, "Knowing How You Know: Children's Understanding of the Sources of Their Knowledge," *Child Development*, LIX (1988): 1366–71.

To illustrate this point, consider the super-*duper*-blindsighter, who is reliable in forming not only first-order beliefs about objects in the blind field, but also second-order beliefs about the unconscious perceptual states on the basis of which his first-order beliefs are formed.[38] Intuitively, his second-order beliefs, just like his first-order beliefs, are no more rational than beliefs formed on the basis of blind guesswork. After all, there is nothing accessible from the subject's own perspective to suggest that they are *not* formed on the basis of guesswork. Unlike the second-order beliefs of the normally sighted, they are not formed on the basis of any conscious perceptual experience. In conclusion, then, the epistemic role of consciousness is what explains the epistemic asymmetry between conscious perceptual experience and unconscious perception in blindsight.

Moreover, the epistemic role of consciousness explains why blindsighted subjects can think *descriptive* thoughts about objects in the blind field but not *demonstrative* thoughts. The crucial point is that they do not satisfy the epistemic constraints on the possession of demonstrative concepts. In particular, their unconscious perception of objects in the blind field does not provide them with immediate justification to form either descriptive or demonstrative beliefs about those objects. This is compatible with thinking descriptive thoughts about those objects, such as the thought that there are objects in the blind field which they cannot see. But given the existence of epistemic constraints on the possession of demonstrative concepts, it excludes the possibility of demonstrative thought.

IV

So far, I have been arguing that conscious perceptual experience plays an essential role in explaining our grasp of demonstrative concepts. I will now consider whether the argument can be generalized in support of the claim that there is also an essential role for conscious perceptual attention. The most straightforward strategy invokes the *limitation thesis*, which says that one experiences an object only if it captures one's attention.[40] If the limitation thesis is true, then we can generalize the argument for the role of conscious perceptual experience simply by substituting "attention" for "experience" throughout. But is there any reason to suppose that the limitation thesis is true?

---

[38] The coherence of this example depends on rejecting higher-order thought theories of consciousness. For example, see David Rosenthal, "A Theory of Consciousness," in Block, Owen Flanagan, and Guven Gazeldere, eds., *The Nature of Consciousness: Philosophical Debates* (Cambridge: MIT, 1997), pp. 729–53.

Some have argued that the limitation thesis is supported by empirical results on *inattentional blindness*. In one salient example, subjects are asked to count the number of times a basketball is passed among a team of players dressed in matching uniforms.[41] During the task, someone dressed in a gorilla suit walks into the midst of the players, beats his chest, and walks off again. When asked later, many subjects fail to report the occurrence of anything unusual. What explains this surprising result? One possible explanation is that our attention is distracted by the task of counting the passes, so we do not *experience* the gorilla. If this explanation is correct, it provides some evidence for the thesis that we experience an object only if we attend to it. However, an alternative explanation is that our attention is distracted, and so we do not *notice* the gorilla, which supports only the weaker thesis that we notice an object only if we attend to it. It is consistent with this weaker thesis that we sometimes experience an object that we fail to notice, which seems sufficient to explain why we do not report it. Therefore, it is not clear that the limitation thesis is supported by the empirical data.

Is there any reason to suppose that the limitation thesis is false? Some have argued, on broadly introspective grounds, that we experience things that fail to capture our attention and therefore escape our notice. For example, Block suggests that I may be dimly aware of the sound of a drill in the street outside, although I do not notice the sound until it stops. However, there is a problem here. How can I know by introspection that I was aware of the drilling sound all along, since it did not capture my attention until it stopped? More generally, how can we have introspective grounds to suppose that we experience things that fail to capture our attention?

In response, one might appeal to the role of memory: I can remember the sound of the drill just a moment ago. Michael Martin argues that memories typically derive from earlier perceptions and that it is by virtue of their link with earlier perceptions that memories provide information about the past.[42] On these grounds, he argues that the contents of memories provide defeasible evidence concerning the contents of earlier perceptions. Thus, if I now remember the sound of drilling just a moment ago, then it is reasonable for me to believe that a moment ago I must have perceived the sound of the drill. But why suppose that the contents of memories provide defeasible evidence concerning the contents of earlier *conscious* perceptions? Why not suppose that my memories are derived from earlier perceptions that are unconscious because unattended?

---

[40] For a more detailed discussion of arguments for and against the limitation thesis, see Smithies, "Attention as Rational-Access Consciousness."

[41] Daniel Simons and Christopher Chabris, "Gorillas in Our Midst," *Perception*, xxviii, 9 (1999): 1059–74.

A key point here is that experiential memory is *Janus-faced*: if I can remember the sound of the drill just a moment ago, then I can also remember what it was like for me to hear the sound of the drill just a moment ago. In the same way, I can remember what it was like to experience the agony of toothache after the pain has subsided. In other words, experiential memory provides information not only about past states of the world, but also about past states of conscious experience of the world. Therefore, experiential memory can provide introspective justification for claims about one's past experiences. To illustrate the point, suppose you were suddenly to lose all background conscious experience that lies outside the scope of your attention. If you can remember what it was like for you just a moment ago, then you will notice how this is strikingly different from what it is like for you now. In other words, you will have introspective grounds to believe that a moment ago you were conscious of much more than was occupying your attention at that time. Therefore, introspection can provide reason to believe that the limitation thesis is false.

If the limitation thesis is false, then the objects of one's experience do not always capture one's attention. In other words, there is a *phenomenal* distinction to be drawn between the attended foreground and the unattended background of one's experience. This raises the question of whether there is a corresponding *epistemic* distinction between the attended foreground and the unattended background of one's experience. To illustrate, consider the Sea of Faces example one more time. Suppose your companion asks you questions about an object in your visual field which is not visually salient or highlighted, so that your visual experience is like a sea of faces. If you can guess reliably, this will come as a complete surprise to you. Intuitively, your epistemic predicament is similar to blindsight insofar as forming beliefs about an unattended object is no more rational than forming beliefs on the basis of blind guesswork. If the object does not engage your attention but rather escapes your notice, then you cannot form immediately justified beliefs about the object. Your beliefs about the object are immediately justified only if your attention is engaged in such a way that you notice the object and its properties.

---

[42] Michael Martin, "Perception, Concepts, and Memory," *Philosophical Review*, CI, 4 (October 1992): 745–63.

How should we characterize the epistemic role of attention? One answer is that attention to an object is necessary for having immediate justification to form beliefs about the object. If this is correct, then we can generalize the argument of the previous section by substituting "attention" for "experience" throughout. However, it is difficult to motivate this account of the epistemic role of attention without assuming the limitation thesis.[43] Why suppose that attention to an object is necessary for having immediate justification to form beliefs about the object? If we say that attention to an object is necessary for experience to represent the object and its properties, then we assume the limitation thesis. But if we assume that the limitation thesis is false, then attention to an object is not necessary for experience to represent the object and its properties, so why should we suppose that attention to the object is necessary for having immediate justification to form beliefs about the object? In other words, why suppose there is any epistemic asymmetry between the attended and unattended aspects of experience?

My response is that even if attention to an object is not necessary for the object and its properties to be represented in one's experience, it is necessary for converting the representational contents of experience into the contents of thought. Here, we may contrast two different aspects of the functional role of attention. First, attention plays a role in modulating the representational contents of experience, such that differences in the direction of attention make for differences in represented detail.[44] Even if the limitation thesis is false, it is still implausible to suppose what Alva Noe has called a snapshot conception of experience, according to which, "You open your eyes and—*presto!*—you enjoy a richly detailed picture-like experience of the world, one that represents the world in sharp focus, uniform detail and high resolution from the center out to the periphery."[45]

---

[43] I owe this objection to Johannes Roessler. Compare his objections to McDowell's conceptualism in Roessler, "Perceptual Attention and the Space of Reasons," in Mole, Smithies, and Wu, eds., *op. cit.*, pp 274–91.

[44] For experimental support, see Marisa Carrasco, Sam Ling, and Sarah Read, "[illegible]tion Alters Appearance," *Nature Neuroscience*, VII (2004): 308–13.

[45] Alva Noe, "Is the Visual World a Grand Illusion?" *Journal of Consciousness Studies*, IX, 5–6 (2002): 1–12, at p. 2.

Calicut University Library CAMPUS

All the same, attention plays a role not only in *modulating* the contents of experience, but also in *formatting* the contents of experience in such a way as to make them accessible for use in conceptual thought. Thus, attention to an object is necessary for converting the contents of experience into the contents of justified belief.[46]

This suggests an alternative account of the epistemic role of attention. On this account, attention to an object is not necessary for *having* immediate justification to form beliefs about the object, but it is necessary for *using* one's immediate justification in forming beliefs about the object that are immediately justified.[47] This reveals a crucial difference between genuine blindness and mere inattentional blindness. If one is blind to an object in the sense that one does not experience the object, then one lacks immediate justification to form beliefs about the object. By contrast, if one is merely inattentively blind to an object, in the sense that one experiences the object but it fails to capture one's attention, then one has immediate justification to form beliefs about the object, but one is unable to use this justification in forming immediately justified beliefs about the object.

To illustrate the contrast between these two accounts of the epistemic role of attention, consider the argument that attention to an object is necessary for being justified in trading on the identity of the object in making inferences from the premises "that is F" and "that is G" to the conclusion "that is F and G."[48] Intuitively, one is not justified in forming beliefs about the identity of a particular object unless one attends to the object in question. Even if one is justified in believing that something is F and that something is G, one is not thereby justified in forming beliefs about which thing is F and which thing is G; in particular, one is not justified in forming beliefs about whether these things are one and the same. In the absence of attention, one's beliefs about such matters are no more justified than beliefs formed on the basis of blind guesswork. This intuitive point is supported by Anne Treisman's experimental work on "illusory conjunctions," which demonstrates that subjects tend to make mistakes about the identity of an object when attention is drawn away.[49] For example, subjects report a red "T" when a green "T" and a red "O" are presented together and attention is distracted.

---

[46] I develop this account of the functional role of attention in Smithies, "Attention as Rational-Access Consciousness." There, I argue that attention is what makes information fully accessible for use in the rational control of action, verbal report, and belief formation.

[47] In the standard terminology, attention is not necessary for *propositional* justification, but it is necessary for *doxastic* justification. One's belief that p is doxastically justified if and only if one *has* propositional justification to believe that p and, moreover, one *uses* one's propositional justification in believing that p on the basis of one's propositional justification to believe that p.

[48] For this argument, see Campbell, *op. cit.*, chapter 5; see also Imogen Dickie, "Visual Attention Fixes Demonstrative Reference by Eliminating Referential Luck," in Mole, Smithies, and Wu, eds., *op. cit.*, pp. 292–322.

Why is attention to an object necessary for being justified in trading on the identity of the object? According to Treisman's feature-integration theory of attention, attention to the location of an object is necessary for binding together the properties represented at that location as properties of a single object. Treisman's theory suggests that attention to an object, or its location, is necessary for trading on the identity of the object because it is necessary for solving the *binding problem*—that is, the problem of binding properties together as properties of a single object. However, it can be misleading to talk about *the* binding problem, since there are multiple binding problems, which arise at multiple levels of psychological reality. In particular, there are distinct binding problems that arise at (i) the level of computational information processing; (ii) the level of perceptual experience; and (iii) the level of conceptual thought. There is empirical evidence to suggest that attention is not necessary for binding at the level of computational information processing.[50] However, it is a further question whether attention is necessary for binding at the level of perceptual experience or the level of conceptual thought.[51]

On my proposal, attention is necessary for solving the binding problem at the level of conceptual thought but not the level of perceptual experience. Attention is not necessary for one's experience to represent properties as bound to a single object. Moreover, attention is not necessary for one's experience to provide one with immediate, defeasible justification to believe that properties are bound to a single object and to make inferences that trade on the identity of the object. Nevertheless, attention is necessary for solving the binding problem at the level of conceptual thought: it is necessary for using one's justification to believe that properties are bound to a single object and to make inferences that trade on the identity of the object. Indeed, this is why attention is necessary for demonstrative thought. Grasping demonstrative concepts does not merely require the presence of immediate justification, but also the ability to use it in forming justified beliefs about an object. So, just as before, we can argue for the role of attention in demonstrative thought on the basis of a premise about the epistemic role of attention together with a further premise about the epistemic conditions necessary for the possession of demonstrative concepts:

---

[49] Treisman and Hilary Schmidt, "Illusory Conjunctions in the Perception of Objects," *Cognitive Psychology*, xiv, 1 (January 1982): 107–41

[50] Patients with unilateral neglect cannot attend to objects in the left hemifield, but representation of objects in the left hemifield primes performance in recognizing objects in the right hemifield. See Anna Berti and Giacomo Rizzolatti, "Visual Processing without Awareness: Evidence from Unilateral Neglect," *Journal of Cognitive Neuroscience*, iv, 4 (1992): 345–51.

[51] For a recent discussion that acknowledges some of these complexities, see Treisman, "Consciousness and Perceptual Binding," in Axel Cleeremans, ed., *The Unity of Consciousness: Binding, Integration, and Dissociation* (New York: Oxford, 2003), pp. 95–113.

The Attention Argument

(1) One has a demonstrative concept of an object *o* only if one has information about *o* which provides immediate, defeasible justification to form beliefs about *o* and which one is able to use in forming immediately justified beliefs about *o*.

(2) One has information about *o* which provides immediate, defeasible justification to form beliefs about *o* and which one is able to use in forming justified beliefs about *o* only if one has conscious perceptual attention to *o*.

(3) One has a demonstrative concept of an object *o* only if one has conscious perceptual attention to *o*.

V

This paper began by asking what perception must be like in order to explain our capacity to think demonstrative thoughts about the world around us. I have argued that perception of an object explains our capacity to think demonstrative thoughts about the object by enabling us to form immediately justified beliefs about the object. Moreover, I have argued that perception of an object enables us to form immediately justified beliefs about an object only if it is conscious and attentive. I therefore conclude that perception must be conscious and attentive in order to play its epistemic role in explaining our capacity for demonstrative thought.

There are further questions to be asked about the nature of conscious and attentive perceptual experience. Is experience a relation to objects and properties in the world around us, or is it a relation to a representational content? Does experience involve the exercise of conceptual abilities, or is it a more primitive, nonconceptual state? My arguments for the epistemic role of experience do not depend on any particular account of what experience is like. However, the epistemic role of experience imposes constraints on an account of the nature of experience, since any such account is answerable to the question: what must experience be like in order to play its epistemic role?

There are also further questions to be asked about the nature of demonstrative thought. The scope of this paper is restricted to demonstrative thought that is based on perception, but we can also think demonstrative thoughts on the basis of memory and testimony. Ultimately, one might hope for a unified account of demonstrative thought that explains what these different varieties have in common. A promising proposal is that any demonstrative concept of an object involves some kind of information link with an object via perception, memory, or testimony, which enables one to form immediately justified beliefs about the object. Moreover, one might expect my arguments for the epistemic role of consciousness and attention to generalize to information links of these various kinds. Suppose one's information link with an object enables one to form immediately justified beliefs about the object only if it is conscious and attended. This yields a more general account of the role of consciousness in demonstrative thought.

Does consciousness play a more general and foundational role in explaining our capacity for conceptual thought? One issue is whether the role of consciousness in demonstrative thought extends to other varieties of information-based thought, including thought that involves the exercise of recognitional capacities or the use of proper names.[52] A further issue is how much of our conceptual thought ultimately depends on demonstrative thought and other kinds of information-based thought.[53] If the dependence is widespread, then this makes a strong case for the classical Russellian thesis that consciousness plays a foundational role in explaining our capacity for conceptual thought.

In *The Problems of Philosophy*, Bertrand Russell writes, "All our knowledge, both knowledge of things and knowledge of truths, rests upon acquaintance as its foundation."[54] Acquaintance, for Russell, is a conscious state of direct and unmediated awareness of things. Meanwhile, the distinction between knowledge of things and knowledge of truths corresponds to the distinction between our conceptual ability to think about things and our epistemic ability to know truths about those things. So, the Russellian thesis is that consciousness plays a dual role in explaining our conceptual ability to think about things and our epistemic ability to know truths about those things. This raises a question about the relationship between these two aspects of the role of consciousness. Indeed, why should we suppose that there is any single thing that unifies these two distinct roles?

---

[52] On recognitional concepts and proper names, see Evans, *op. cit.*, chapters 8, 9, and 11.

[53] Compare the argument from massive reduplication in Peter Strawson, *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen, 1959), chapter 1.

[54] Bertrand Russell, *The Problems of Philosophy* (Oxford: University Press, 1912).

DOES PROBABILITY THEORY REFUTE COHERENTISM?

The central claims of the coherence theory of justification are (i) that beliefs are justified by virtue of coherence, and (ii) that no belief has foundational justification. Over the last decade, some philosophers have sought to use probability theory both to understand and to evaluate this theory. Out of this recent work have emerged grave doubts about the tenability of coherentist doctrine. When one tries to translate claims about epistemic justification into claims about probability, the coherence theory appears to fly in the face of standard mathematical probability theory. This, in turn, would seem to provide powerful grounds for siding with foundationalists against the champions of coherence.

This, as I shall argue, is an illusion. Much as one might wish to see this particular philosophical debate resolved through mathematical proof, the resources of probability theory have proven inadequate to the task of adjudicating between foundationalism and coherentism. None claim that probability theory refutes foundationalism; unfortunately, neither does it refute coherentism.

In the following, after discussing the core commitments of coherentism, I review three challenges to the coherence theory based on theorems of probability that the coherentist allegedly runs afoul of. Two of these challenges have appeared previously in the literature. In each case, I argue that the challenge misconstrues the coherentist's commitments, and that, understood properly, the coherentist's commitments can be satisfied by some coherent probability distributions.

I. THE HEART OF COHERENTISM

Coherentism arose out of dissatisfaction with foundationalism as a solution to the epistemic regress problem. The characteristic claim of foundationalism that many find problematic is that certain beliefs have foundational justification, where this is understood as a form of justification that does not require support by other justified beliefs. Many have found this sort of justification mysterious. Objections to the notion include (a) that any belief is arbitrary to the extent that it is not supported by justifiably believed reasons and (b) that foundationalists cannot explain why putatively foundational beliefs are likely to be true.[1] This is not the place to evaluate these objections;

---

[1] See Peter Klein, "Human Knowledge and the Infinite Regress of Reasons," in James E. Tomberlin, ed., *Philosophical Perspectives 13: Epistemology* (Cambridge: Blackwell, 1999),

0022-362X/11/0801/035–054

they serve, however, to call to mind the motivations behind the coherentist alternative—it is a theory that, at least in intention, avoids the central pitfalls of foundationalism.

The coherentist's first fundamental commitment, then, is to the following thesis:

> *Nonfoundationalism*: No belief has any justification independent of its relations to other beliefs that support or cohere with it.

The coherentist thus rejects not merely the notion of foundational knowledge, but that of a belief's having even the smallest degree of foundational justification. This commitment emerges directly from the imperative to avoid the supposed flaws of foundationalism—the objections from the alleged arbitrariness of merely foundational beliefs and from the alleged impossibility of accounting for their likely truth are not merely objections to taking beliefs that lack inferential support to be *highly* justified. They are objections to taking such beliefs to be justified to any extent. Coherentists therefore seek to construct all epistemic justification without recourse even to the weakest sort of foundations.[2]

Some recent authors appear to regard Nonfoundationalism as negotiable, as they reject a probabilistic constraint meant to embody Nonfoundationalism, leaving nothing in its place. Perhaps their reasoning is that since a Nonfoundationalist theory is obviously untenable in the light of probability theory, we may, to render the discussion of coherentism interesting, discard Nonfoundationalism as a part of coherentist doctrine.[3] This, however, does not speak to the central motivations of coherentism. Because Nonfoundationalism emerges straightforwardly from those motivations, it is not a dispensible component of the coherentist doctrine. If no defensible Nonfoundationalist theory can be constructed, then coherentism should simply be abandoned. In fact, however, we shall see in sections II and III that a Nonfoundationalist coherentism can be represented probabilistically, without falling afoul of any theorems of probability theory.

---

pp. 297–325, at p. 297; Laurence BonJour, *The Structure of Empirical Knowledge* (Cambridge. Harvard, 1985), pp. 30–33.

[2] Particularly clear on this point is BonJour, *op. cit.*, pp 26–30, 148, who finds what he calls 'weak foundationalism' even less plausible than stronger versions of foundationalism.

[3] Luc Bovens and Stephan Hartmann, *Bayesian Epistemology* (New York: Oxford, 2003), pp. 14–15 gesture in this direction, rejecting the constraint I label below as 'Strong Nonfoundationalism'. Bovens and Erik Olsson, "Coherentism, Reliability and Bayesian Networks," *Mind*, n. s., CIX, 436 (2000): 685–719, at pp. 711–12 give similar hints.

Having rejected foundationalist accounts of justification, the coherentist stands in need of an alternative source of epistemic justification. This leads to the second major tenet of coherentism:

> *Coherence Justification*: Some beliefs are justified by virtue of coherence.

The sort of coherence at issue is commonly taken to involve explanatory and support relations (both deductive and probabilistic) among beliefs.[4] Fortunately, the details regarding the analysis of coherence need not be filled in to evaluate the probabilistic challenges to the coherence theory with which we shall be concerned in the following sections.

The intuition behind Coherence Justification is brought forth in the following sort of scenario. A detective interviews two witnesses to a crime, giving neither witness the chance to confer with the other. If the witnesses tell divergent stories, then little can be concluded. But suppose that the witnesses independently tell essentially the same story, agreeing on very specific and unanticipated details about the crime. For instance, suppose they both agree that the culprit had spiked green hair. The detective would then have strong reason for thinking that the witnesses were telling the truth; otherwise, such specific agreement would be highly improbable. This would be true even if the detective had no prior reason to trust in the witnesses' reliability or honesty.[5]

The scenario can be generalized: rather than two witnesses to a crime, suppose simply that one has two information sources, or belief-forming mechanisms, with no initial justified opinion about how trustworthy either source is. These sources might be vision and hearing, for example, or distinct lines of reasoning based on independent observations. If these two sources "agree" on a set of specific details, then one has reason to conclude that the propositions on which they agree are in fact true, since otherwise the agreement would have to be regarded as a very improbable coincidence.

This scenario is not intended to establish the truth of coherentism. It is intended merely to explain how Coherence Justification and Nonfoundationalism might both be true. With no antecedent justified belief about the reliability of either information source, it is plausible that one would have no justification for accepting the deliverances of either source considered by itself. Foundationalists may of course dispute this, but for now our aim is to clarify the coherentist's intuitions. Despite the lack of justification for trusting in a single information source, when the two sources converge justification apparently emerges from their combined testimony. This seems to demonstrate at least the coherence of the coherence theory of justification.

---

[4] See BonJour, *op. cit.*, pp. 95–99; Keith Lehrer, *Knowledge* (New York: Oxford, 1974), pp. 163–65; Paul Thagard, *Coherence in Thought and Action* (Cambridge: MIT, 2000), pp. 43, 53.

[5] C. I. Lewis, *An Analysis of Knowledge and Valuation* (La Salle, IL: Open Court, 1946), p. 346, and BonJour, *op. cit.*, pp. 147–48 discuss this sort of case. The spiked green hair example is Catherine Elgin's, "Non-Foundationalist Epistemology: Holism, Coherence, and Tenability," in Matthias Steup and Ernest Sosa, eds., *Contemporary Debates in Epistemology* (Malden, MA: Blackwell, 2005), pp. 156–67, at p. 157.

The key step in this line of thinking is a piece of probabilistic reasoning: we should believe that the witnesses are correct, because if they were not telling the truth the coherence of their testimonies would be improbable. It is therefore natural to analyze the example using probability theory. In the following two sections, we scrutinize the probabilistic credentials of the proposed illustration of coherentism.

II. THE WEAK FOUNDATIONALIST CHALLENGE

In the witness scenario we have described, the two witnesses independently report some proposition, *A*. Neither witness has any independent credibility; that is, neither witness's report by itself would provide one with any justification for believing *A*. Nevertheless, the combined reports are held to justify one to some degree in believing *A*. To represent these conditions probabilistically, let W_1 be the proposition that the first witness asserts *A*, and let W_2 be the proposition that the second witness asserts *A*. We may lay down the following conditions:

(1) *Conditional Independence*: $P(W_2|W_1,A) = P(W_2|A)$
$P(W_2|W_1,{\sim}A) = P(W_2|{\sim}A)$
(2) *Strong Nonfoundationalism*: $P(A|W_1) = P(A)$
$P(A|W_2) = P(A)$
(3) *Coherence Justification*: $P(A|W_1,W_2) > P(A)$

Conditional Independence is intended to represent the stipulation that the witnesses make their reports independently, with no opportunity to influence one another. This seems to imply that, when the truth-value of *A* is fixed, what one witness asserts has no probabilistic effect on what the other witness asserts.

I have labeled the second pair of probabilistic conditions 'Strong Nonfoundationalism' for reasons that will emerge in section III. These conditions are meant to represent the stipulation that neither witness's testimony by itself would give one any justification for believing *A*. This being so, it seems that the probability of *A* should be unchanged by either witness's individual assertion.

Lastly, condition (3) represents a minimal reading of Coherence Justification, namely, that the witnesses' combined, convergent testimonies should provide at least some justification for believing what they assert.

This may be interpreted as meaning that the probability of A is raised by the conjoined witness testimonies.

Now, the first major challenge to the coherence theory is this: conditions (1), (2), and (3) are jointly incompatible. No coherent probability distribution over any three propositions makes all of those mathematical relations hold.[6] If the coherence theory of justification is genuinely committed to all of those claims, then the coherence theory is incoherent.

But the coherence theory is not committed to all of those claims. In particular, the coherentist may reject the assumption of Conditional Independence. This assumption is motivated by the desire to avoid the sort of case in which two witnesses fudge their testimonies into agreement—for instance, the second witness asks the first witness what the first witness is going to tell us, whereupon the second witness decides to tell us the same thing. In such a situation, the agreement of the witnesses would do nothing to enhance the credibility of their story. Conditional Independence rules out this sort of situation.[7]

However, while Conditional Independence *suffices* to rule out witness collaboration, it is not *required* for ruling out such collaboration. Collaboration can be ruled out by the weaker assumption that, if a given proposition is *false*, one witness's asserting it does not make the other witness more likely to assert it, that is, $P(W_2|W_1,{\sim}A) \leq P(W_2|{\sim}A)$.[8] Moreover, the underlying motivation for stipulating independence was simply to articulate conditions favorable to coherence's being epistemically valuable. We wanted to avoid a scenario in which witnesses fudge their testimonies into agreement, precisely because in such a scenario coherence obviously does not enhance credibility. If this was the underlying rationale for assuming independence of witness testimonies, then it stands to reason that we should if possible avoid interpreting the independence assumption in such a way that it *undermines* the ability of coherence to enhance credibility.

It turns out that, given Strong Nonfoundationalism, the following condition is necessary and sufficient for Coherence Justification:

---

[6] This theorem is proved in my "Probability and Coherence Justification," *Southern Journal of Philosophy*, xxxv, 4 (Winter 1997): 463–72, where I mistakenly took it to refute coherentism, and in Olsson's *Against Coherence: Truth, Probability, and Justification* (New York: Oxford, 2005), pp. 68–69, 218. Throughout, I suppress mention of assumptions regarding nonextreme probabilities. For instance, it should be understood that in this theorem, $P(W_1,W_2)$, $P(A)$, and $P({\sim}A)$ are all nonzero.

[7] Olsson, *Against Coherence*, pp. 113–15; Bovens and Hartmann, *Bayesian Epistemology*, pp. 15–16; Huemer, *op. cit.*, pp. 466–67.

[8] L. Jonathan Cohen, *The Probable and the Provable* (New York: Oxford, 1977), pp. 101–13 advances this requirement, along with the natural parallel requirement that, given the *truth* of a proposition, one witness's asserting it does not make the other *less* likely to assert it, that is, $P(W_2|W_1,A) \geq P(W_2|A)$.

$P(W_2|W_1,A) > P(W_2|W_1,{\sim}A)$.[9] This is the condition that, given that the first witness asserts A, the other witness is more likely to concur if A is true than if A is false. This would be true if A itself bears favorably on the reliability of the witnesses; for example, if information gathered by the sense of sight suggests that one's senses are generally reliable, then information gathered by the sense of hearing is more likely to corroborate those visual observations if the visual observations were accurate than if they were not. This probabilistic condition also rules out the sort of collaboration in which the second witness repeats whatever the first witness says regardless of whether it is true. The coherentist should therefore adopt the probabilistic condition that $P(W_2|W_1,A) > P(W_2|W_1,{\sim}A)$, in place of Conditional Independence.

III. THE WEAK FOUNDATIONALIST CHALLENGE, ROUND TWO

The weak foundationalist's objection to coherentism is that one cannot combine Nonfoundationalism with Coherence Justification. We have just rejected one argument for this claim. But there remains a more troubling argument relying on weaker assumptions.[10] The conjunction of Strong Nonfoundationalism and Coherence Justification entails the following thesis:

Negative Coherence Justification: $P(A|{\sim}W_1,{\sim}W_2) > P(A)$

This amounts to the claim that if both witnesses "agree" in *not* asserting A, this fact *supports* A. A witness may fail to assert A either by remaining silent on the question of whether A or by denying A. Presumably, no coherentist will happily claim that if the witnesses each either remain silent or deny A, one thereby has some justification for accepting A.

The mathematical derivation of Negative Coherence Justification, given Strong Nonfoundationalism and Coherence Justification, appears in Appendix II. Here is the intuitive idea. We know from Strong Nonfoundationalism that W_1 has no effect on the probability of A. Now, W_1 may be divided into two sub-possibilities: $(W_1 \,\&\, W_2)$ and $(W_1 \,\&\, {\sim}W_2)$. According to Coherence Justification, the first of these sub-possibilities *raises* the probability of A. Therefore, the other one must *lower* the probability of A, if W_1 on the whole is to be neutral with respect to A. So $P(A|W_1 \,\&\, {\sim}W_2) < P(A)$. Next, notice that since W_2 is probabilistically irrelevant to A, ${\sim}W_2$ is equally irrelevant to A. And ${\sim}W_2$ may also be divided into two sub-possibilities: $(W_1 \,\&\, {\sim}W_2)$ and

---

[9] A proof appears as Appendix I to this paper. I discuss the theorem further in my "Weak Bayesian Coherentism," *Synthese*, CLVII, 3 (August 2007): 337–46

[10] I thank a referee for this JOURNAL for pointing out a theorem related to the objection discussed in the text.

$(\sim W_1 \,\&\, \sim W_2)$. We have just seen that the first of these sub-possibilities lowers the probability of A. Therefore, the other sub-possibility must raise the probability of A, if $\sim W_2$ on the whole is to be neutral with respect to A. Hence, we have the thesis of Negative Coherence Justification: $P(A|\sim W_1,\sim W_2) > P(A)$.

This unacceptable result follows from Coherence Justification and Strong Nonfoundationalism. Coherence Justification is a non-negotiable commitment of the coherence theory. I have also argued that Nonfoundationalism is non-negotiable. So the coherentist seems to be at the end of his rope.

I see exactly one reasonable response for the coherentist. This is to question the probabilistic interpretation of Nonfoundationalism. Nonfoundationalism denies that any belief has any foundational justification. In the witness scenario under discussion, this is represented by the condition that an individual witness's testimony by itself has no effect on the probability of A. But what exactly does *that* mean? Consider two probabilistic interpretations of the claim that the first witness's testimony by itself has no effect on the probability of A:

Strong Nonfoundationalism: $P(A|W_1) = P(A)$

Weak Nonfoundationalism: $P(A|W_1,\sim W_2) = P(A)$

The former equation tells us that, given just the information that the first witness has asserted A, there is no reason to believe A. The latter equation tells us that, given just the information that the first witness has asserted A *and the other witness has not*, there is no reason to believe A. Which is a better interpretation of Nonfoundationalism?

The coherentist holds that in a certain possible situation, one lacks justification for believing A. We represent the absence of such justification as a lack of impact on the probability of A, that is, as A's retaining its initial probability, $P(A)$. What is the relevant situation? Presumably, it is the situation of one's having received exactly one witness report affirming A. This is not to be confused with the situation of one's having received *at least* one witness report affirming A. The former situation obtains precisely when either $(W_1 \,\&\, \sim W_2)$ or $(\sim W_1 \,\&\, W_2)$ is true. Assuming that one has symmetrical attitudes towards the two witnesses, either of those alternatives has the same impact on the probability of A. Thus, the coherentist's claim should be understood as that $(W_1 \,\&\, \sim W_2)$ does not affect the probability of A. The situation of one's having received *at least* one witness report affirming A, on the other hand, obtains when either $(W_1 \,\&\, \sim W_2)$ or $(\sim W_1 \,\&\, W_2)$ or $(W_1 \,\&\, W_2)$ is true. The third of these alternatives is the situation in which one has coherent witness testimonies; this alternative therefore must not be included as a determinate of the scenario in which one has "a single witness's testimony," for purposes of stating the Nonfoundationalist condition. So the Nonfoundationalist claim should be understood as one about $P(A|W_1,{\sim}W_2)$ and $P(A|W_2,{\sim}W_1)$, rather than about $P(A|W_1)$.

We can arrive at this conclusion in a slightly different way. Suppose you have received a report that *A* from the first witness and no such report from the second witness. You have no other information relevant to *A*. The probability you now assign to *A* should be the probability of *A* conditional on your total relevant evidence. Your total relevant evidence is not W_1, since you also know something relevantly different and logically stronger than W_1: you know $(W_1 \mathbin{\&} {\sim}W_2)$, since you have received the first witness report that *A* and have not received a second witness report that *A*. Thus, you must conditionalize on $(W_1 \mathbin{\&} {\sim}W_2)$. The coherentist's claim that in this situation you have no justification for *A* therefore amounts to the claim that $P(A|W_1,{\sim}W_2)$ is the same as the prior probability of *A*.

One might object that you do not know ${\sim}W_2$ when you merely fail to observe that W_2, because you do not know that the second witness will not assert *A* at some future time. In response, there are at least two ways of interpreting 'W_1' and 'W_2': one could interpret 'W_1' as referring to the proposition that the first witness asserts *A* at some time or other, or one could interpret 'W_1' as referring to the proposition that the first witness has asserted *A* at or before the present time (that is, the time at which the probability of *A* is to be assessed); and similarly for 'W_2'. The latter interpretation is preferable, because it better reflects your total evidence: at time *t*, if you have received the first witness's report that *A* and have not received the second witness's report that *A*, then you know something logically stronger than that the first witness reports that *A* at some time or other; you know that the first witness reported that *A* during the time period that you observed the first witness at or before *t*, and that the second witness did not report that *A* during this time period. Thus, you should conditionalize on $(W_1 \mathbin{\&} {\sim}W_2)$, where both conjuncts are interpreted as describing what happened during the observational time period up to *t*.[11]

---

[11] Another objection is that $(W_1 \mathbin{\&} {\sim}W_2)$ is not one's total evidence, because one must also know something logically stronger than $(W_1 \mathbin{\&} {\sim}W_2)$· either one knows that W_1 and the second witness denied *A*, or one knows that W_1 and the second witness remained silent with regard to whether *A*. However, the coherentist would take either of these more specific situations as leaving one with no justification for affirming or denying *A*; thus, their disjunction, which is $(W_1 \mathbin{\&} {\sim}W_2)$, has no effect on the probability of *A*, as claimed in the text.

Once we replace Strong Nonfoundationalism with Weak Nonfoundationalism, the anti-coherentist arguments discussed in this section and the preceding section dissolve. It is easily proven that the coherentist's two core commitments of Coherence Justification and Weak Nonfoundationalism are compatible. The following is a possible probability distribution (see Figure 1):

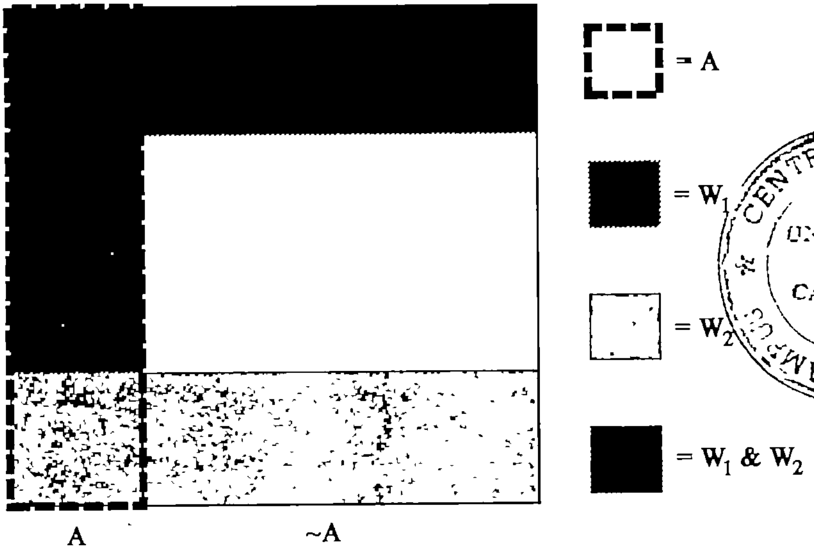


Figure 1: A probability distribution satisfying Weak Nonfoundationalism and Coherence Justification. Probabilities are represented by areas on the diagram, with the area of the whole square equal to 1.

Possibility	Probability
A, W_1, W_2	2/16
$A, W_1, \sim W_2$	1/16
$A, \sim W_1, W_2$	1/16
$A, \sim W_1, \sim W_2$	0
$\sim A, W_1, W_2$	0
$\sim A, W_1, \sim W_2$	3/16
$\sim A, \sim W_1, W_2$	3/16
$\sim A, \sim W_1, \sim W_2$	6/16

The above table shows the probabilities assigned to each maximally specific possibility; for instance, (A & W_1 & W_2) has a probability of 2/16. It is easily verified that Weak Nonfoundationalism is satisfied, since $P(A) = 1/4 = P(A|W_1,{\sim}W_2) = P(A|{\sim}W_1,W_2)$. Coherence Justification is perfectly satisfied, since $P(A|W_1,W_2) = 1$. And Negative Coherence Justification is avoided, since $P(A|{\sim}W_1,{\sim}W_2) = 0$. Though this last equation may seem overly strong, it is not, at any rate, an anti-coherentist condition—we may think of it as the condition that, if both witnesses agree in *not* asserting A, A is conclusively refuted. The witnesses' conjoint failure to assert A may be construed as a sort of coherence; thus, the present probability distribution portrays this rather minimal sort of coherence as having a dramatic effect. I do not contend that this probability distribution is realistic or that its existence vindicates the coherence theory. Rather, the point is that the existence of this probability distribution demonstrates that probability theory is not hostile to the coherence theory in the manner previously suspected.

IV. THE TRUTH-CONDUCIVENESS CHALLENGE

The final probabilistic challenge to coherentism begins from the assumption that the coherentist needs a *truth-conducive* measure of coherence. Suppose that a number of witnesses have each asserted some proposition. The witness reports may cohere more or less with each other. Coherence is said to be truth-conducive if and only if, *ceteris paribus*, the more coherent the set of witness reports is, the more likely it is that all the reports are true. A *truth-conducive measure* of coherence is a way of measuring the degree of coherence of a set of reports, such that coherence is guaranteed to be truth-conducive.

Recent work has revealed that, subject to assumptions to be discussed below, no truth-conducive measure of coherence is possible.[12] Given any measure of coherence, it is always possible to find a case in which greater coherence for a set of witness reports leads to a lower overall probability of truth in the light of the witnesses' testimony. Erik Olsson takes this theorem to show that coherence, as conceived by the coherentist, does not exist. Less dramatically, one might argue that while there may be such a thing as coherence, it cannot satisfy the epistemological aspirations that coherentists have pinned on it.[13]

For the sake of argument, I assume here that the coherentist accepts the commitment to the general truth-conduciveness of coherence.[14] Nevertheless, I shall argue that the coherentist can survive the challenge from the impossibility theorems of Olsson and Luc Bovens and Stephan Hartmann.

---

[12] See Olsson, *Against Coherence*, chapter 7; Bovens and Hartmann, *Bayesian Epistemology*, chapter 1, though Bovens and Hartmann believe the coherence theory is salvageable.

[13] Olsson, *op. cit.*, pp. 137–39.

[14] As I argue in "Weak Bayesian Coherentism," it is questionable whether this is a genuine commitment of the coherence theory.

IV.1. Assumptions of the Impossibility Theorems. The proofs of the impossibility of a truth-conducive measure of coherence rely on the following two assumptions:

> *Content Determination*: The degree of coherence of a set of witness reports is a function of the probabilities of the reported propositions and their boolean combinations.
>
> *Conditional Independence*: Once the truth-values of the reported propositions are fixed, what one witness says is probabilistically independent of what any other witnesses say.

The notion of truth-conduciveness is formalized as follows. Let $\mathbf{A} = \{A_1, ..., A_n\}$ and $\mathbf{B} = \{B_1, ..., B_n\}$ be two sets of propositions. Suppose we have two sets of witnesses, $\{w_1, ..., w_n\}$ and $\{v_1, ..., v_n\}$. Let W_{A_1} be the proposition that witness w_1 asserts A_1; V_{B_1}, the proposition that v_1 asserts B_1; W_{A_2}, the proposition that w_2 asserts A_2; and so on. Now, suppose $\mathbf{A}$ is more coherent than $\mathbf{B}$, and all other epistemically relevant factors are equal. To say that coherence is truth-conducive is to say that in any situation of this kind, $P(A_1, ..., A_n | W_{A_1}, ..., W_{A_n}) > P(B_1, ..., B_n | V_{B_1}, ..., V_{B_n})$.

What is meant in this context by the stipulation that other epistemically relevant factors are equal? What things must be held equal to satisfy this *ceteris paribus* clause? Consider three possible interpretations:

> *CP1*: Other things are equal provided that $P(A_1 | W_{A_1}) = P(B_1 | V_{B_1})$, ..., and $P(A_n | W_{A_n}) = P(B_n | V_{B_n})$.
>
> *CP2*: Other things are equal if:
> (i) $P(A_1 | W_{A_1}) = P(B_1 | V_{B_1})$, ..., and $P(A_n | W_{A_n}) = P(B_n | V_{B_n})$; *and*
> (ii) $P(A_1, ..., A_n) = P(B_1, ..., B_n)$.
>
> *CP3*: Other things are equal *only* if $P(A_1) \times ... \times P(A_n) = P(B_1) \times . . \times P(B_n)$.

CP1 only requires holding fixed *witness reliability*, the probability of a witness report being correct: reliability levels for the witness reports in **A** should be the same as those for the corresponding members of **B**. This is Olsson's interpretation of the *ceteris paribus* clause.[15]

CP2 requires holding fixed both witness reliability and *joint prior probability*: the initial probability of the conjunction of the members of **A** should be the same as that of the conjunction of the members of **B**. This is Bovens and Hartmann's interpretation.[16]

---

[15] Olsson, *op. cit.*, pp. 119–23.

[16] Bovens and Hartmann, *op. cit.*, pp. 11–12.

CP3 requires holding fixed the *total individual strength* of the two sets of reported propositions; that is, the product of the prior probabilities of the members of **A** should be the same as the product of the prior probabilities of the members of **B**. This is Tomoji Shogenji's condition.[17]

As I have stated them, CP1 and CP2 each state sufficient conditions for "other things" to be equal, whereas CP3 states only a necessary condition. CP1 entails CP2, whereas CP3 contradicts both CP1 and CP2. CP3 claims that there is a condition overlooked by CP1 and CP2 that must be held fixed when we consider whether coherence is truth-conducive.

The challenge to the truth-conduciveness of coherence requires three distinct assumptions: Content Determination, Conditional Independence, and either CP1 or CP2. If any of these assumptions is rejected, the impossibility theorem can no longer be derived. I have questioned Conditional Independence in section II above. In the following two subsections, I call into question Content Determination, CP1, and CP2.

IV.2. Is Coherence Content-Determined? Coherence is content-determined just in case the degree of coherence of a set of witness reports is determined entirely by the probability distribution over the propositions asserted by the witnesses.

There are grounds for doubting that coherence is content-determined in this sense. The Content Determination assumption plays a crucial role in a conflict of intuitions regarding coherence that has recently appeared in the literature. Branden Fitelson finds it intuitively clear that coherence should be maximal for a set of logically equivalent propositions.[18] After all, coherentists traditionally hold that coherence is mainly a matter of mutual support among beliefs. For a set of logically equivalent propositions, mutual support would seem to be at its maximum: any subset of the propositions in question entails any other subset.

Igor Douven and Wouter Meijs, however, cultivate a contrary intuition: imagine interviewing a witness who, after telling a not very coherent story, returns to the beginning and produces first one paraphrase and then another of the same story. For each statement he makes, the witness produces two other, logically equivalent statements. Intuitively, he has not thereby made his testimony more coherent. We surely should say this if we want coherence to be truth-conducive, since the witness has not rendered his story more plausible merely by restating it.[19]

---

[17] Tomoji Shogenji, "Is Coherence Truth Conducive?" *Analysis*, LIX, 4 (October 1999). 338–45, at p. 342.

[18] Branden Fitelson, "A Probabilistic Theory of Coherence," *Analysis*, LXIII, 3 (July 2003): 194–99, at p. 194.

If we treat coherence as a content-determined property, then Fitelson's intuition and Douven and Meijs's intuition are irreconcilable—it cannot be that a set of logically equivalent propositions both is and is not highly coherent. But if we allow the coherence of a set of reports to depend on more than just its content, we can reconcile the apparently conflicting intuitions by supposing that Fitelson and Douven and Meijs have in mind different sorts of reporting contexts. Contrast the following two scenarios:

> *Case 1*: Seventeen witnesses to a crime all report that the culprit had spiked green hair.
>
> *Case 2*: Only Jon reports that the culprit had spiked green hair, but he says it seventeen times.

Intuitively, Case 1 is an impressive instance of coherence, due mainly to the number of agreeing witnesses, but Case 2 is utterly unimpressive as a case of coherence. Jon's seventeenfold assertion is not significantly more coherent than a single assertion by Jon to the effect that the culprit had spiked green hair. Yet the Content Determination assumption would force us to assign the same degree of coherence in Cases 1 and 2—each case merely exhibits the degree of coherence, whatever it may be, of the set $\{G\}$, where G = the proposition that the culprit had spiked green hair.[20] If, on the other hand, we reject Content Determination, we may include features of the testimonial context other than the report contents—features such as whether the reports originated from the same source or distinct sources—among the determinants of the coherence of a set of reports.[21] This makes room for embracing Douven and Meijs's intuition about their case, while simultaneously embracing Fitelson's intuition in cases of reports by distinct witnesses.

A similar response applies to an objection to the coherence theory due to Richard Fumerton. Fumerton observes that if one believes any two propositions, P and Q, and one has at least a modest knowledge of logic and is moderately reflective, one will also believe the material biconditional $(P \leftrightarrow Q)$. This gives one a set of believed propositions, $\{P, Q, (P \leftrightarrow Q)\}$, in which each member is entailed by the others. This would seem to guarantee a high degree of coherence, yet as Fumerton observes, this putative coherence is too easily obtained to confer any significant degree of justification on the propositions involved. If this set of beliefs is highly coherent, then anyone with a modest knowledge of logic can (and perhaps already does, at least dispositionally) hold a highly coherent system of beliefs.[22]

---

[19] Igor Douven and Wouter Meijs, "Measuring Coherence," *Synthese*, CLVI, 3 (April 2007): 405–25, at p. 417.

[20] Alternately, if we make use of ordered n-tuples rather than sets (perhaps to distinguish the case of a single witness from that of multiple agreeing witnesses), Cases 1 and 2 each would exhibit the degree of coherence, whatever it may be, of the ordered 17-tuple $\langle G, G, G, G, G, G, G, G, G, G, G, G, G, G, G, G, G\rangle$.

[21] Shogenji, "Reply to Akiba on the Probabilistic Measure of Coherence," *Analysis*, LXI, 2 (April 2001): 147–50, at pp. 149–50 makes a related suggestion, though his proposal concerns the individuation of beliefs

Fumerton's problem may be resolved by allowing coherence to depend on features of a belief system other than its propositional content. In this case, the fact that the third belief in the system was acquired solely by inference from the other two, rather than from some independent belief source, might be held to undermine the would-be coherence of the beliefs in the system with each other, just as, in Douven and Meijs's example, the fact that only one witness produces the multiple concurrent reports undermines the would-be coherence of those reports.

A further reason for rejecting Content Determination is that it seems to rule out coherence as a source of empirical justification. If coherence is determined by the prior probabilities of the propositions that our witnesses assert, then the degree of coherence of a set of propositions is knowable *a priori*; hence, any justification that coherence confers on them would seem to be *a priori*. If—as the coherentist maintains—all justification is generated by coherence, and—as the Content Determination assumption assures us—the degree of coherence of a set of reports depends only on the initial probabilities of their propositional contents, then our actually *receiving* a collection of reports should have no effect on our justification for believing the reported propositions—the coherence of those propositions has not increased, and nothing else is relevant to justification. One might of course maintain that in a testimonial situation, our background beliefs, themselves justified by coherence, dictate that the receiving of the reports is evidentially relevant. However, this sort of reply would not be available in the general case of the justification of our belief system; the coherentist would still be forced to treat one's belief system as *a priori* justified. For this reason, coherentists are well advised to treat coherence as a property depending on more than just belief contents.

---

[22] Richard Fumerton, *Metaepistemology and Skepticism* (Lanham, MD: Rowman & Littlefield, 1995), pp. 145–47.

IV.3. When Are Other Things Equal? The interpretation of the *ceteris paribus* clause in the definition of truth-conduciveness turns out to be crucial to assessing the challenge to the coherence theory. If we follow Shogenji in interpreting the *ceteris paribus* clause along the lines of CP3, then there may be truth-conducive measures of coherence, such as Shogenji's own proposed measure:

$$C_1(\mathrm{A}) = \frac{\mathrm{P}(A_1 \,\&\ldots\&\, A_n)}{\mathrm{P}(A_1)\times\ldots\times P(A_n)}$$

where **A** is a set of propositions whose members are $A_1, \ldots, A_n$. As Shogenji observes, C_1 is trivially truth-conducive in the sense that if we hold fixed the total individual strength of **A** (the denominator in the above ratio), the greater the value of C_1 is, the more likely it is that all the members of **A** are true.[23] While this does not establish that C_1 is truth-conducive in the sense defined earlier—namely, that the greater C_1 is, the higher is the conjoint probability of the members of **A** *conditional on* the occurrence of the respective witness reports—neither do the impossibility results of Olsson and Bovens and Hartmann rule this out.

The impossibility results are obtained only by rejecting total individual strength as one of the "other things" that must be held equal when we examine the effect of coherence on probability of truth. Here is one reason to side with Shogenji on this issue: when investigating the contribution of coherence to likelihood of truth, we must hold fixed factors that are independent of coherence. It is plausible that the individual prior probabilities, $\mathrm{P}(A_1)$, $\mathrm{P}(A_2)$, and so on, are independent of coherence, since coherence is typically characterized as entirely a matter of relationships among propositions. Thus, $\mathrm{P}(A_1|A_2)$ is partly determinative of (or perhaps determined by) the coherence between A_1 and A_2, but it is hard to see how the initial probability just of A_1 might be a matter of coherence among the A_i. Therefore, it seems that "other things are equal" only when individual prior probabilities are held fixed, that is, $\mathrm{P}(A_1) = \mathrm{P}(B_1), \ldots,$ and $\mathrm{P}(A_n) = \mathrm{P}(B_n)$. And this entails that $\mathrm{P}(A_1) \times \ldots \times \mathrm{P}(A_n) = \mathrm{P}(B_1) \times \ldots \times \mathrm{P}(B_n)$.

Why, then, does Olsson reject CP3? He begins by observing that if we want to investigate the truth-conduciveness of some property F, we should hold fixed epistemically relevant factors that are independent of F, but we should not hold fixed factors that depend on F or on which F depends. But on Shogenji's measure, the degree of coherence of an information set apparently depends on its total individual strength. One reason for believing this is that C_1-coherence is *defined* partly in terms of total individual strength. Another reason, advanced by Olsson, is that a specification of the total individual strength of a set of propositions *constrains* the value of C_1: while C_1 has no general upper bound, if the total individual strength of a belief set is given as x, C_1 acquires an upper bound of $1/x$.[24]

---

[23] Shogenji, "Is Coherence Truth Conducive?"

The latter argument suggests the principle that a given feature, F, must be held fixed if and only if specifying the value of F places no constraints on the degree of coherence of a set of witness reports. But this criterion yields inconsistent results: as discussed, it yields the result that if coherence is measured by C_1, then total individual strength need not be held fixed. However, when coherence is measured by C_1, each individual prior probability, $P(A_i)$, by itself fails to constrain the degree of coherence of a set of reports.[25] So every individual prior probability must be held fixed—yet somehow their product need not be.

Even if we could accept Olsson's principle, it would remain open to a coherentist to adopt a tailor-made measure of coherence such as the following:

$$C_2(\mathbf{A}) = \frac{\tanh^{-1}[P(A_1 \,\&\, \ldots \,\&\, A_n)]}{P(A_1) \times \ldots \times P(A_n)}$$

The inverse hyperbolic tangent function is chosen here solely for its property of being a monotonically increasing function that maps $[0,1)$ onto $[0,\infty)$. Any function with that property will deliver the desired result: fixing total individual strength no longer constrains the value of C, provided that n is allowed to vary. Suppose that the total individual strength of $\mathbf{A}$ is x (where $0 < x < 1$). $C_2(\mathbf{A})$ may still fall anywhere between zero and infinity.[26] C_2 thus makes total individual strength independent of coherence on Olsson's criterion of independence. If we hold total individual strength fixed, $P(A_1 \,\&\, \ldots \,\&\, A_n)$ also increases monotonically with $C_2(\mathbf{A})$, so C_2 is truth-conducive.

What of the argument that, on Shogenji's view, total individual strength is not independent of coherence because coherence is partly defined in terms of total individual strength? This argument seems to rest on a mistake. Let I be the total individual strength of an information set;

---

[24] Olsson, *op. cit.*, pp 117–18, 122.

[25] If $P(A_1)$ is given as x, the minimum possible value of C_1 remains 0 (achieved, for example, when A_1 is incompatible with A_2). There remains no maximum value of C_1, provided $n > 2$. To increase C_1 without limit, assume $P(A_2) = P(A_3) = P(A_1 \,\&\, \ldots \,\&\, A_n)$, and let $P(A_2)$ approach zero. Then C_1, which will be at least $1/P(A_2)$, approaches infinity.

[26] $C_2(\mathbf{A})$'s lower bound is zero, since $P(A_1 \,\&\, .. \,\&\, A_n)$'s lower bound is zero and $\tanh^{-1}(0) = 0$. To increase C_2 without limit, let the A_i be logically equivalent so that $P(A_1 \,\&\, .. \,\&\, A_n) = P(A_1) = \sqrt[n]{x}$, and let n increase without bound. Then $C_2(\mathbf{A}) = \frac{\tanh^{-1}\sqrt[n]{x}}{x}$ approaches infinity as n increases.

J, the joint prior probability of its members; and *C*, its degree of coherence. Shogenji's proposal to measure coherence by the ratio *J*/*I* should not mislead one into thinking that on his view coherence depends on *J* and *I*. Rather, his conception seems to be that *J* is determined by two factors, *I* and *C*, in such a way that $I \times C = J$. Though the relation can be restated as "$C = J/I$," this should not obscure the fact that *I* and *C* are the independent variables, *J* the dependent variable. Coherence does not *consist of* a favorable *J*/*I* ratio; rather, coherence consists of the mutual support relations among propositions that are *responsible* for the difference between *I* and *J*.[17]

Either C_1 or C_2, then, appears to provide a suitable counterexample to the impossibility theorems of Olsson and Bovens and Hartmann.

V. CONCLUSION: A MODEST VICTORY FOR COHERENTISM

All of the recent probabilistic challenges to the coherence theory rest on questionable assumptions. The weak foundationalist challenge of section II argues that if there is no foundational justification for information from a single source, then no coherence justification emerges for convergent information from two or more sources. This challenge depends on the questionable assumptions of Strong Nonfoundationalism and Conditional Independence. The challenge of section III claims that if coherence provides justification in the absence of foundational justification, then the *absence* of testimonies supporting a given proposition also provides justification for that proposition. But this challenge depends on the questionable assumption of Strong Nonfoundationalism. Finally, the challenge of section IV claims that there is no generally truth-conducive measure of coherence. This challenge depends on the questionable assumptions of Conditional Independence and Content Determination, as well as a questionable interpretation of the phrase 'other things being equal'.

The coherentist may reject Conditional Independence, assuming instead that one information source is more likely to agree with a true report than a false report from another information source. He may reject Content Determination, arguing that the coherence of a set of reports depends not just on their contents, but also on their sources. He may replace Strong Nonfoundationalism with Weak Nonfoundationalism, postulating that a report from one information source, conjoined with the absence of any convergent report from any other source, fails to generate justification for the content of that report. Finally, the coherentist may reasonably require that the product of the individual prior probabilities of a set of reported propositions be held fixed when assessing the probabilistic effects of coherence. Given this understanding of the coherentist's commitments, the coherentist escapes both the weak foundationalist challenges and the truth-conduciveness challenge.

---

[17] Shogenji, "Is Coherence Truth Conducive?" p. 342.

None of this is to say that coherentism is a plausible or correct epistemological theory. While it is mathematically possible for testimonies that would be useless in isolation to prove valuable in combination, it is equally possible for individually useless testimonies to prove useless in combination, even when their contents agree. My own view remains that foundationalism is the most plausible account of the structure of our actual justified beliefs. My purpose here, however, is not to assess the relative merits of coherentism and foundationalism in general. Rather, my purpose has been to assess the impact of recent work in probability theory on the foundationalism/coherentism debate. Unfortunately, recent efforts to resolve that debate by appeal to probability theory have been unsuccessful, due in large part to the failure to motivate sufficient constraints on the concept of coherence and on the coherence theory generally. In the absence of significantly stronger constraints of these kinds than so far have been justified, the probability calculus cannot tell us whether the coherence theory is true.

MICHAEL HUEMER

University of Colorado, Boulder

APPENDIX I: CONDITIONS UNDER WHICH STRONG NONFOUNDATIONALISM AND COHERENCE JUSTIFICATION ARE CONJOINTLY TRUE

Assume: (a) $P(A|W_1) = P(A)$
(b) $P(A, W_1, W_2) > 0$
(c) $P(\sim A) > 0$

To prove: $P(A|W_1, W_2) > P(A)$ iff $P(W_2|W_1, A) > P(W_2|W_1, \sim A)$

Proof:

(1) $P(A|W_1,W_2) = \frac{P(A)\ P(W_1,W_2|A)}{P(A)\ P(W_1,W_2|A) + P(\sim A)\ P(W_1,W_2|\sim A)}$ Bayes' Theorem

(2) $P(A|W_1,W_2) = \frac{P(A)\ P(W_1|A)\ P(W_2|W_1,A)}{P(A)\ P(W_1|A)\ P(W_2|W_1,A) + P(\sim A)\ P(W_1|\sim A)\ P(W_2|W_1,\sim A)}$ (1)

(3) $P(W_1|A) = P(W_1)$ Assumptions (a), (b)

(4) $P(W_1|\sim A) = P(W_1)$ Assumptions (a), (c)

(5) $P(A|W_1,W_2) = \frac{P(A)\ P(W_1)\ P(W_2|W_1,A)}{P(A)\ P(W_1)\ P(W_2|W_1,A) + P(\sim A) \cdot P(W_1)\ P(W_2|W_1,\sim A)}$ (2), (3), (4)

(6) $P(A|W_1,W_2) = \frac{P(A)\ P(W_2|W_1,A)}{P(A) \cdot P(W_2|W_1,A) + P(\sim A) \cdot P(W_2|W_1,\sim A)}$ Simplifying (5)

(7) $P(A|W_1,W_2) > P(A)$ iff $\frac{P(A) \cdot P(W_2|W_1,A)}{P(A)\ P(W_2|W_1,A) + P(\sim A)\ P(W_2|W_1,\sim A)} > P(A)$ (6)

(8) $P(A|W_1,W_2) > P(A)$ iff
$P(W_2|W_1,A) - P(A)\ P(W_2|W_1,A) > P(\sim A) \cdot P(W_2|W_1,\sim A)$ Algebraic manipulation of (7)

(9) $P(A \mid W_1,W_2) > P(A)$ iff $P(W_2 \mid W_1,A)[1 - P(A)] > [1 - P(A)]P(W_2 \mid W_1,\sim A)$	(8)
(10) $P(A \mid W_1,W_2) > P(A)$ iff $P(W_2 \mid W_1,A) > P(W_2 \mid W_1,\sim A)$	(9), assumption (c)

APPENDIX II: DERIVATION OF NEGATIVE COHERENCE JUSTIFICATION

First Lemma: For any propositions p and q with nonzero probabilities, $\frac{P(q \mid p)}{P(q)} = \frac{P(p \mid q)}{P(p)}$

Proof of First Lemma:

(1) $P(p,q) = P(q,p)$	Logically equivalent propositions
(2) $P(p) \times P(q \mid p) = P(q) \times P(p \mid q)$	(1)
(3) $\frac{P(q \mid p)}{P(q)} = \frac{P(p \mid q)}{P(p)}$	(2)

Second Lemma: For any propositions p, q, r with nonzero probabilities, if $P(p \mid q) = P(p)$, then $P(q,r \mid p) - P(q,r) = P(q,\sim r) - P(q,\sim r \mid p)$.

Proof of Second Lemma:

(1) $P(p \mid q) = P(p)$	Assumption
(2) q is logically equivalent to $([q,r] \vee [q,\sim r])$	Propositional logic
(3) $P(q) = P([q,r] \vee [q,\sim r])$	(2)
(4) $P(q) = P(q,r) + P(q,\sim r)$	(3)
(5) $P(q \mid p) = P([q,r] \vee [q,\sim r] \mid p)$	(3)
(6) $P(q \mid p) = P(q,r \mid p) + P(q,\sim r \mid p)$	(5)
(7) $P(q \mid p) = P(q)$	(1), First Lemma
(8) $P(q) = P(q,r \mid p) + P(q,\sim r \mid p)$	(6), (7)
(9) $P(q,r \mid p) + P(q,\sim r \mid p) = P(q,r) + P(q,\sim r)$	(4), (8)
(10) $P(q,r \mid p) - P(q,r) = P(q,\sim r) - P(q,\sim r \mid p)$	(9)
(11) If $P(p \mid q) = P(p)$, then $P(q,r \mid p) - P(q,r) = P(q,\sim r) - P(q,\sim r \mid p)$	(1)–(10), conditional proof

Derivation of Negative Coherence Justification (assuming $P(W_1,W_2,A)$, $P(\sim W_1,\sim W_2) > 0$):

(1) $P(A \mid W_1) = P(A \mid W_2) = P(A)$	Assumption (Strong Nonfoundationalism)
(2) $P(A \mid W_1,W_2) > P(A)$	Assumption (Coherence Justification)
(3) $P(W_1,W_2 \mid A) - P(W_1,W_2) = P(W_1,\sim W_2) - P(W_1,\sim W_2 \mid A)$	(1), Second Lemma (letting $p = A$, $q = W_1$, $r = W_2$)

(4)	$P(W_1,W_2 \mid A) > P(W_1,W_2)$	(2), First Lemma (letting $p = [W_1 \& W_2]$, $q = A$)
(5)	$P(W_1,\sim W_2) > P(W_1,\sim W_2 \mid A)$	(3), (4)
(6)	$P(\sim W_2,W_1) > P(\sim W_2,W_1 \mid A)$	(5)
(7)	$P(A \mid \sim W_2) = P(A)$	(1)
(8)	$P(\sim W_2,W_1 \mid A) - P(\sim W_2,W_1) = P(\sim W_2,\sim W_1) - P(\sim W_2,\sim W_1 \mid A)$	(7), Second Lemma (letting $p = A$, $q = \sim W_2$, $r = W_1$)
(9)	$P(\sim W_2,\sim W_1 \mid A) > P(\sim W_2,\sim W_1)$	(6), (8)
(10)	$P(A \mid \sim W_2,\sim W_1) > P(A)$	(9), First Lemma (letting $p = A$, $q = [\sim W_1 \& \sim W_2]$)
(11)	If $P(A \mid W_1) = P(A \mid W_2) = P(A)$ and $P(A \mid W_1,W_2) > P(A)$, then $P(A \mid \sim W_2,\sim W_1) > P(A)$	(1)–(10), conditional proof

CAMBRIDGE

OUTSTANDING SCHOLARSHIP *in* PHILOSOPHY

Civil Religion
A Dialogue in the History of Political Philosophy
RONALD BEINER
$95.00 Hb 978-0-521-50636-6
$34.99 Pb 978-0-521-73843-9
448 pp

Politics in Dark Times
Encounters with Hannah Arendt
SEYLA BENHABIB
$90.00 Hb 978-0-521-76370-7
$28.99 Pb 978-0-521-12722-6
408 pp

The Sociology of Theodor Adorno
MATTHIAS BENZER
$95.00 Hb 978-1-107-00009-4
278 pp

Cognitive Science
An Introduction to the Science of the Mind
JOSE LUIS BERMUDEZ
$120.00 Hb 978-0-521-88200-2
$60.00 Pb 978-0-521-70837-1
516 pp

The Cambridge Companion to Ancient Scepticism
RICHARD BETT
Cambridge Companions to Philosophy
$95.00 Hb 978-0-521-87476-2
$32.99 Pb 978-0-521-69754-5
392 pp

The Practice of Global Citizenship
LUIS CABRERA
$90.00 Hb 978-0-521-19936-0
$31.99 Pb 978-0-521-12810-0
328 pp

Ancient Self-Refutation
The Logic and History of the Self-Refutation Argument from Democritus to Augustine
LUCA CASTAGNOLI
$99.00 Hb 978-0-521-89631-3
414 pp

Visions of Discovery
New Light on Physics, Cosmology, and Consciousness
RAYMOND Y. CHIAO,
MARVIN L. COHEN,
ANTHONY J. LEGGETT,
WILLIAM D. PHILLIPS, *and*
CHARLES L. HARPER, JR.
$85.00 Hb 978-0-521-88239-2
826 pp

The Monochord in Ancient Greek Harmonic Science
DAVID CREESE
Cambridge Classical Studies
$110.00 Hb 978-0-521-84324-9
426 pp

The Dark Side of Creativity
DAVID H. CROPLEY,
ARTHUR J. CROPLEY,
JAMES C. KAUFMAN, *and*
MARK A. RUNCO
$95.00 Hb 978-0-521-19171-5
$36.99 Pb 978-0-521-13960-1
404 pp

Capitalism, For and Against
A Feminist Debate
ANN CUDD *and*
NANCY HOLMSTROM
$99.00 Hb 978-0-521-11407-3
$29.99 Pb 978-0-521-13211-4
360 pp

Euthanasia and Physician-Assisted Suicide
GERALD DWORKIN,
R. G. FREY, *and*
SISSELA BOK
For and Against
$87.00 Hb 978-0-521-58246-9
$38.99 Pb 978-0-521-58789-1.
160 pp

Towards Discursive Education
Philosophy, Technology, and Modern Education
CHRISTINA E. ERNELING
$85.00 Hb 978-0-521-19474-7
$31.99 Pb 978-0-521-14402-5
210 pp

John Locke's Politics of Moral Consensus
GREG FORSTER
$32.99 Pb 978-0-521-18118-1
332 pp

The Texts of Early Greek Philosophy
The Complete Fragments and Selected Testimonies of the Major Presocratics
DANIEL W. GRAHAM
$180.00 2 Hb bks
978-0-521-84591-5
$99.00 2 Pb bks
978-0-521-60842-8 1,168 pp

Dialogical Self Theory
Positioning and Counter-Positioning in a Globalizing Society
HUBERT HERMANS *and*
AGNIESZKA HERMANS-KONOPKA
$99.00 Hb 978-0-521-76526-8
404 pp

Fichte's Social and Political Philosophy
Property and Virtue
DAVID JAMES
Modern European Philosophy
$80.00 Hb 978-1-107-00155-8
234 pp

Making the Political
Founding and Action in the Political Theory of Zhang Shizhao
LEIGH K. JENCO
$95.00 Hb 978-0-521-76060-7
296 pp

Prices subject to change

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

from CAMBRIDGE UNIVERSITY PRESS

Psychoanalysis in a New Light
GUNNAR KARLSSON
$95 00 Hb 978-0-521-19805-9
$34 99 Pb 978-0-521-12244-3
232 pp

Religion and the Political Imagination
IRA KATZNELSON *and*
GARETH STEDMAN JONES
$95 00 Hb 978-0-521-76654-8
$29 99 Pb 978-0-521-14734-7
394 pp

The Cambridge Handbook of Creativity
JAMES C KAUFMAN *and*
ROBERT J STERNBERG
Cambridge Handbooks in Psychology
$135 00 Hb 978-0-521-51366-1
$65 00 Pb 978-0-521-73025-9
508 pp

Listening for Utopia in Ernst Bloch's Musical Philosophy
BENJAMIN M. KORSTVEDT
$95 00 Hb 978-0-521-89615-3
234 pp

The Ironic Defense of Socrates
Plato's Apology
DAVID M. LEIBOWITZ
$80 00 Hb 978-0-521-19479-2
204 pp

Being, Nature, and Life in Aristotle
Essays in Honor of Allan Gotthelf
JAMES G LENNOX *and*
ROBERT BOLTON
$99 00 Hb 978-0-521-76844-3
306 pp

Prices subject to change

Explanation and Teleology in Aristotle's Science of Nature
MARISKA LEUNISSEN
$85 00 Hb 978-0-521-19774-8
264 pp

Global Justice and Due Process
LARRY MAY
$95 00 Hb 978-0-521-76272-4
$32 99 Pb 978-0-521-15235-8
260 pp.

Machiavellian Democracy
JOHN P McCORMICK
$90 00 Hb 978-0-521-82390-6
$27 99 Pb. 978-0-521-53090-3
264 pp

Now in Paperback!
The Philosophy of Gottlob Frege
RICHARD L MENDELSOHN
$29 99 Pb 978-0-521-18119-8
246 pp

Thinking the Unconscious
Nineteenth-Century German Thought
ANGUS NICHOLLS *and*
MARTIN LIEBSCHER
$95 00. Hb 978-0-521-89753-2
344 pp

Ancient Models of Mind
Studies in Human and Divine Rationality
ANDREA NIGHTINGALE *and*
DAVID SEDLEY
$95 00 Hb 978-0-521-11355-7
260 pp

Plato on Music, Soul and Body
FRANCESCO PELOSI
Translated by SOPHIE HENDERSON
$95 00 Hb 978-0-521-76045-4
236 pp

Displacement by Development
Ethics, Rights and Responsibilities
PETER PENZ,
JAY DRYDYK *and*
PABLO S BOSE
$90 00 Hb 978-0-521-19882-0
$31 99 Pb 978-0-521-12464-5
358 pp

The Origins of Aesthetic Thought in Ancient Greece
Matter, Sensation, and Experience
JAMES I PORTER
$145 00 Hb 978-0-521-84180-1
624 pp

Third Edition!
Practical Ethics
PETER SINGER
$90 00 Hb 978-0-521-88141-8
$31 99 Pb 978-0-521-70768-8
352 pp

Fifth Edition!
The Psychologist's Companion
A Guide to Writing Scientific Papers for Students and Researchers
ROBERT J STERNBERG *and*
KARIN STERNBERG
$95 00 Hb 978-0-521-19571-3
$32 99 Pb 978-0-521-14482-7
376 pp

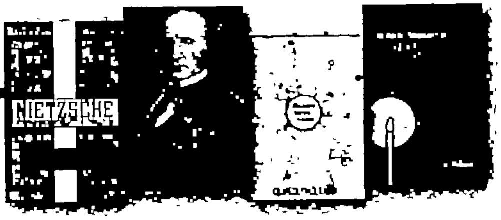
Now in Paperback!
The Transformation of Learning
Advances in Cultural-Historical Activity Theory
BERT VAN OERS,
WIM WARDEKKER,
ED ELBERS, *and*
RENÉ VAN DER VEER
$32 99 Pb 978-0-521-15698-1
418 pp

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

New from

Oxford



The Unity of Consciousness
TIM BAYNE
Tim Bayne examines the idea that a human being can have only a single stream of consciousness at any one point in time.
2010 256 pp. $55.00

Nietzsche and the Ancient Skeptical Tradition
JESSICA N. BERRY
This work presents a portrait of Nietzsche as the skeptic par excellence in the modern period
2010 248 pp $65.00

John Stuart Mill and the *Art of Life*
Edited by BEN EGGLESTON, DALE E. MILLER, and DAVID WEINSTEIN
The *Art of Life* is John Stuart Mill's name for his account of practical reason. In this volume eleven leading scholars elucidate this fundamental, but widely neglected, element of Mill's thought.
2010 320 pp $74.00

The Best Things in Life
A Guide to What Really Matters
THOMAS HURKA
An illuminating look at the best things in life—what they are, how to get them, and how they help make our lives more meaningful
(Philosophy in Action)
2010 208 pp $18.95

Meaning, Mind, and Matter
Philosophical Essays
ERNIE LEPORE and BARRY LOEWER
Two leading philosophers draw together their work on meaning and mind.
2010 280 pp $65.00

The Philosophy of Generative Linguistics
PETER LUDLOW
Peter Ludlow presents the first book on the philosophy of generative linguistics
2010 240 pp $55.00

Attention Is Cognitive Unison
An Essay in Philosophical Psychology
CHRISTOPHER MOLE
The first philosophical theory of attention for many years, and the first to draw extensively on research from the cognitive sciences
(Philosophy of Mind Series)
2010 224 pp. $49.95

The Moral Dimensions of Human Rights
CARL WELLMAN
Wellman brings clarity to the ways in which differing views on human rights are represented in contexts like municipal law, constitutional law, international law, and humanitarian and human rights law
2010 248 pp $74.00

New in **Paperback**

Supersizing the Mind
Embodiment, Action, and Cognitive Extension
ANDY CLARK
(Philosophy of Mind Series)
2010 320 pp Paperback $19.95

Our Stories
Essays on Life, Death, and Free Will
JOHN MARTIN FISCHER
2010 192 pp. Paperback $24.95

Following the Rules
Practical Reasoning and Deontic Constraint
JOSEPH HEATH
2010 354 pp. Paperback $29.95

How Terrorism Is Wrong
Morality and Political Violence
VIRGINIA HELD
2010 216 pp Paperback $17.95

Making Amends
Atonement in Morality, Law, and Politics
LINDA RADZIK
2010 256 pp Paperback $29.95

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply only in the US
To order or for more information, visit our website at www.oup.com/us

Jacques Rancière
Key Concepts
Edited by Jean-Philippe Deranty
Acumen Publishing
978-1-84465-233-4 $24.95 paper
978-1-84465-232-7 $75.00 cloth

"An accessible but sophisticated introduction to the extraordinarily rich conceptual apparatus of Rancière."
–Andrew Schaap, University of Exeter

Immanuel Kant
Key Concepts
Edited by Will Dudley
and Kristina Engelhard
Acumen Publishing
978-1-84465-239-6 $24.95 paper
978-1-84465-238-9 $75.00 cloth

This collection from top contributors unpacks Kant's notoriously dense and complex ideas and presents them in a form manageable for the beginning student.

Jürgen Habermas
Key Concepts
Edited by Barbara Fultner
Acumen Publishing
978-1-84465-237-2 $24.95 paper
978-1-84465-236-5 $75.00 cloth

An invaluable resource for both novice readers of Habermas and those interested in a more refined understanding of particular aspects of his work.

McGILL-QUEEN'S UNIVERSITY PRESS
www.mqup.ca

THE JOURNAL OF
PHILOSOPHY

CENTRAL LIBRARY
UNIVERSITY
OF
CALCUTTA

VOLUME CVIII, NUMBER 2
FEBRUARY 2011

page
55 *Imagination Is Where the Action Is* — Neil Van Leeuwen
78 *Self-Consciousness and Immunity* — Timothy Lane and Caleb Liang
100 *There Is No Argument that the Mind Extends* — Sam Coleman
109 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2011 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 2, FEBRUARY 2011

IMAGINATION IS WHERE THE ACTION IS*

Two forms of pretending, if properly understood, put a strain on prevalent views about human imagination in philosophy and cognitive science. If we recast our image of imagination, we can accommodate them. The two phenomena I wish to introduce are *semi-pretense* and *pretense layering.*

(For clarity, I always use "pretend" and cognates to refer to bodily behavior; "imagine" and cognates refer to mental activity and structures.)

Semi-pretense: Jennifer and Leroy are at a pool on a summer day, watching other children jump off the high-dive. One clumsy child almost trips off the end of the board and falls into the water at an awkward angle with arms waving. "Two," says Leroy, holding up two fingers with a frown. "I give him a one," says Jennifer, holding up a finger. Next a young woman springs off the board, glides in a swan dive, and enters the water with hardly a splash. "Ten," they both shout, holding up ten fingers each.

Are Jennifer and Leroy *pretending* to be judges at a diving competition, or are they *acting plainly*, simply evaluating the dives for themselves?

The question excludes a middle answer. There is a continuum of cases. At one end (full pretense) Jennifer and Leroy wear polo shirts, speak in officious tones, and hold up placards with numbers. At the other end (plain action), the two just say things like, "that dive was pretty good," with no trappings of a diving judge. The example described is in between, so it is semi-pretense. (If it seems like plain action to you, just move one step over on the continuum.)

---

* For helpful exchanges and comments, I would like to thank Johan van Bentham, Andrew Cortens, Daniel Dennett, Dagfinn Føllesdal, Tamar Gendler, Wes Holliday, Justin Junge, Sean Kelly, Krista Lawlor, Yousri Marzouki, Carlos Montemayor, Sarah Paul, John Perry, Kenneth Taylor, Joseph Ulatowski, and the Harvard Philosophical Psychology Lab. I am also grateful to Gregory Currie and Stephen Stich for email exchanges in which they clarified their positions. Finally, I owe thanks to the Mellon Foundation and the Center for Humanities at Tufts for their support.

0022-362X/11/0802/055–077

Pretense layering: In the film "Rebel Without a Cause," James Dean and Natalie Wood play (pretend to be) high school students, Jim and Judy, who begin dating. One scene is in a deserted mansion, where their two characters pretend to be married adults who consider buying a house. "Should we rent or are we in a buying mood, dear?" "You decide, darling. Remember our budget." Thus, there was more than one layer: Dean and Wood pretended to be high school students who pretend to be adults. In principle, such *pretense layering* can go on for arbitrarily many layers; in practice, there are rarely more than two or three.[1]

These phenomena give theorists a clear directive: identify the imaginative structures responsible for them and other forms of pretense. I take up that directive. To advance our understanding of imagination and pretense, I defend this thesis:

> *Active Imagination Thesis*: There exists a form of imagining that is a continuously updated *forward model* of action in the world, in which action possibilities are constructed in relation to a manifold of largely perceptual representations that can be veridical, nonveridical, or mixed—where *mixed* is a usual and important case.

A "forward model" is an internal representation of motor commands that anticipates the consequences of those commands on bodily motion. The "nonveridical" perceptual representations are basically *mental imagery*. I specify "nonveridical" because some mental imagery is veridical, for example, an accurate visual memory.[2] What is radical about this view, in relation to other views I consider, is the idea that (nonveridical) mental imagery can be *integrated* into the perceptual field and that *this* form of imagining delivers the objects we relate to in constructing pretense action, such as make-believe.[3] My defense of this thesis has three sides.

(1) Competing theories fail to account for important mental and behavioral phenomena, such as semi-pretense and layering. They have other problems, as well.
(2) The theory of imagination I offer accounts for the behavioral phenomena of semi-pretense and layering, as well as other phenomena.
(3) The theory of imagination I offer coheres with empirical findings in cognitive neuroscience.

---

[1] Herb Clark introduces the term "layering" in chapter 12 of *Using Language* (New York: Cambridge, 1996) pp. 353–86 Clark is inspired by Erving Goffman, as am I; see Goffman, *Frame Analysis: An Essay on the Organization of Experience* (Boston: Northeastern, 1986). For philosophical discussion of layering, see Kendall Walton's *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (Cambridge: Harvard, 1990).

[2] "Veridical" basically means *treated as accurate by the agent's cognitive system*. How to flesh out that idea further is an issue I take up in my (unpublished) "What Is a Judgment?"

[3] My theory is radical in relation to some academic views. But the phenomena were at some level familiar all along. I focus here on pretense for play, but my theory extends also to pretense for deception.

My view of imagination—if it is correct—has implications beyond pretense. In the course of this paper, I suggest various ways in which imagination is central to producing action generally.[4]

Sections I and II are critical, displaying the inadequacy of two current theories: (i) the Nichols and Stich theory, which I call *the conditional belief account*, and (ii) what I call *the replacement account*, advocated in somewhat different forms by Velleman, Currie and Ravenscroft, and Doggett and Egan.[5] But the critiques of these competitors lead to constructive suggestions. And section III presents my positive theory: the *integrated imagination theory*. I call it this to emphasize that the structure of representations that leads to pretense (and other) action integrates imagining and veridical representation in a variety of ways. Section IV summarizes the implications of my theory for mind and action.

I. THE CONDITIONAL BELIEF ACCOUNT

Shaun Nichols and Stephen Stich augment the standard desire-belief explanation of action with the "Possible World Box" (PWB), which is meant to house imaginings. The desire-belief explanation of action is that desires and beliefs jointly cause and rationalize actions that will make the contents of the desires true, if the contents of the beliefs are true.[6] (For example: desiring to find a cup of coffee, I go to where I believe the coffee shop is.) How, for Nichols and Stich, does the PWB augment the standard desire-belief explanation of action?

> ...as a possible world description is unfolding in the PWB, the pretender comes to have beliefs of the form:
>
> > If it were the case that p, then it would (or might) be the case that q_1 & q_2 & ... & q_n
>
> where p is the pretence premiss and q_1 & q_2 & ... & q_n are the representations in the PWB. These beliefs, along with the desire to behave in a way

---

[4] Some of what I say may seem to depend on the assumption that Andy Clark criticizes, namely, the assumption of "experience-based control" (in "Visual Experience and Motor Action: Are the Bonds Too Tight?" *Philosophical Review*, CX, 4 (October 2001). 495–519. So I wish to caution the reader that not all forms of imagining I posit are necessarily part of *conscious* experience, although surely some are. Sorting out which are conscious must be postponed until after the theory is stated in representationalist terms.

[5] All references to these authors refer to: Shaun Nichols and Stephen Stich, *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds* (New York: Oxford, 2003); J. David Velleman, "On the Aim of Belief," in *The Possibility of Practical Reason* (New York: Oxford, 2000), pp. 244–81; Gregory Currie and Ian Ravenscroft, *Recreative Minds: Imagination in Philosophy and Psychology* (New York: Oxford, 2002); and Tyler Doggett and Andy Egan, "Wanting Things You Don't Want: The Case for an Imaginative Analogue of Desires," *Philosopher's Imprint*, VII, 9 (December 2007): 1–17.

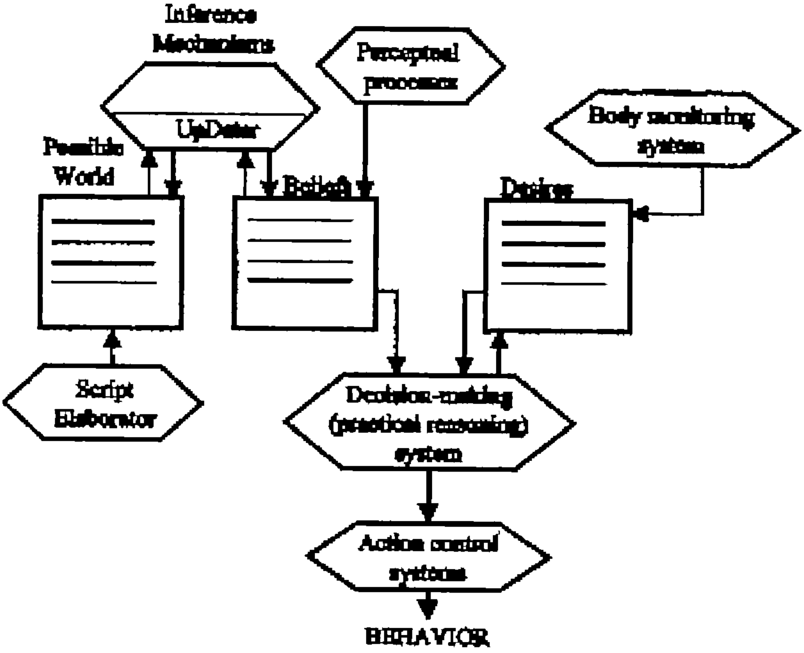
[6] Donald Davidson revived this explanation form in his "Actions, Reasons, and Causes," this JOURNAL, LV, 23 (November 7, 1963): 685–700.

> that is similar to the way one would behave if *p* were the case, lead to the pretence behaviour....[7]

In short, what one imagines generates conditional beliefs, and pretense occurs when one desires to behave as if the antecedent of the conditional belief were true.

Here is how their theory would explain pretending to be a thief. The pretense premise represented in my PWB is *I am a thief.* That premise is elaborated upon in the PWB by various processes, which results in representations with contents such as *I carry a gun into a bank and point it at the teller.* What goes on in the PWB then causes me, the imaginer, to have a conditional belief of the form *if I were a thief, I would carry a gun into a bank and point it at the teller.* Then the bodily pretense, such as making a gun shape with my hand and saying "Stick 'em up!" is caused by the conditional belief and a desire to behave as if the antecedent of the conditional were true. The desire's content would be *that I behave as if I were a thief.*

Importantly, what one imagines on this theory does not have a direct role in causing action. Rather, the effect of imagining on action is mediated by beliefs. The diagram, or "boxology," that Nichols and Stich offer exhibits this mediation:[8]



---

[7] Nichols and Stich, *op. cit.*, pp. 37–38.

[8] *Ibid.*, p. 36.

There is no direct arrow from the PWB (on the left) to the Decision-making system (center). Rather one's imaginings only influence action when mediated by conditional beliefs. That is why I call this the *conditional belief account*. Let me put the point about mediation as starkly as I can: on the Nichols and Stich theory, the pretender's bodily actions would be exactly the same, even if she had no imaginings, as long as the relevant conditional beliefs were still in place.

There are three major shortcomings of this theory. First, it cannot without absurd consequences explain pretense of outlandish behaviors. Second, the claims of the theory devolve into incoherence if one tries to use it to explain *pretense layering*. Third, the theory cannot make sense of the fact that people can *choose* pretense behaviors from a variety of options they think up. These points are not merely negative, however, since understanding them gives rise to constructive suggestions for a better theory of imagining.

Criticism 1 (outlandish pretense): I once saw a comedic improvisation in which two actors, supposed to be in a duel, spontaneously pretended to mount pterodactyls and begin jousting with one another.[9] According to the conditional belief account, explaining this pretense requires positing the following in the Belief Box of either improviser: *if I were in a duel, I would mount a pterodactyl and joust*. But this posit is absurd. No one believes such a thing. So the conditional belief account falters at explaining pretense of outlandish behavior. Nor is this an isolated example. A student who gets a bad grade on her test may, in private, pretend she is shooting the professor with a gun: "Bang, take that!" What would her conditional belief be? Presumably: *if I had a gun, I would shoot my professor*. But surely it is not a requirement that she believe this in order to engage in such pretending. In fact, it is perfectly common for people to pretend to do things that they do *not* believe they would do under any circumstances. Sometimes that is the point.[10]

Criticism 2 (layering): Often in pretense layering, the assumption about who or what one is in one layer is inconsistent (or at least incoherent) with the assumption about who or what one is in the second layer. This can be seen to some extent in the example of James Dean, who is a high school student in one layer and a married adult in another layer. But examples can be starker. Suppose I pretend to be a child who pretends to be a dinosaur on the hunt. What would the conditional belief be in such pretense? Presumably something like: *if I were a child, then (if I were a dinosaur, I would hunt)*. But there is no way to make sense of this belief, since the two antecedents do not cohere at all. If someone claimed to believe that, I would not even understand the claim. Thus, the conditional belief account makes incoherent posits when applied to pretense layering.[11]

---

[9] The group was Stanford Improvisers, otherwise known as SImps.

[10] Criticism of the Nichols and Stich view along these lines also appears in Doggett and Egan, *op. cit.*

Criticism 3 (choice): Sometimes pretend play is utterly spontaneous, but sometimes we choose to pretend and, while pretending, choose between a number of different actions we can perform. In pretending to be a doctor, for example, I can choose between pretending to refuse to perform a surgery and pretending to perform the surgery. How is the conditional belief account to explain the possibility of this sort of choice? Presumably, the pretense premise in the PWB is *I am a doctor*. Then both further contents in the PWB must be represented: *I perform the surgery* and *I refuse to perform the surgery*. At this point, Nichols and Stich must say that I come to have a conditional belief with one of these contents represented in the consequent (whichever it is I actually do). If I end up choosing to pretend to perform the surgery, it must have been because I had the conditional belief *if I were a doctor, I would perform the surgery*; if I end up choosing to pretend not to, it must have been because I had the belief *if I were a doctor, I would refuse to perform the surgery*. But that gets it wrong. When I choose among pretense actions, I do not thereby force one or the other of two conditional beliefs on myself. First, that contradicts the very persuasive idea that coming to have beliefs is *not* a matter of choice.[12] Second, it involves a superfluous and unlikely posit: a belief in something without any evidence whatsoever, the contrary of which could just as well have (by choice!) been believed.

* * *

All three criticisms show that the conditional belief account is forced to posit bizarre, incoherent, or simply unlikely *beliefs* in order to explain ordinary pretense. The account was forced into this position because it posits the Belief Box in between the PWB and the systems that produce behavior. Thus, it is probably a mistake to posit an architecture in which beliefs causally mediate the influence of imagining on pretense action, where this means that imaginings influence behavior only by influencing beliefs. The constructive suggestion, therefore, is that imagining produces pretense directly, without needing the mediation of belief.

---

[11] It is tempting to say that Nichols and Stich can appeal here to the *concept* of PRETEND (*if I were a child, I would PRETEND to be a hunting dinosaur*). But Nichols and Stich disavow appeal to the concept of PRETEND shortly after the quoted passage, since they oppose the way that Alan Leslie invokes that concept. For a snapshot of the debate, see Ori Friedman and Leslie, "The Conceptual Underpinnings of Pretense: Pretending Is Not 'Behaving-As-If'," *Cognition*, cv, 1 (October 2007): 103–24 Also, Stich (personal communication) holds that the passage I quote is meant as a general account of pretense.

[12] The *locus classicus* of this view is Bernard Williams, "Deciding to Believe," in *Problems of the Self* (New York. Cambridge, 2007), pp. 136–51.

It is, of course, undeniable that beliefs in some way influence how and what we pretend. My belief that doctors perform surgery, for example, in part caused me to imagine doing or refusing surgery when I pretended to be a doctor. This shows that beliefs do have a role in imagining and pretense, but Nichols and Stich got it the wrong way around. They said that the contents of imagining bring about conditional beliefs, which then cause pretense behavior. I invert that: beliefs have influence on what one imagines; then imagining primarily influences what one pretends. Furthermore, I think one's beliefs and other cognitive attitudes (as well as conative attitudes) *comment on* and *constrain* what one actively imagines.[13] Beliefs may even have veto power over imagination's ability to yield pretense or other behaviors, but beliefs do not stand between imagining and action.

II. THE REPLACEMENT ACCOUNT

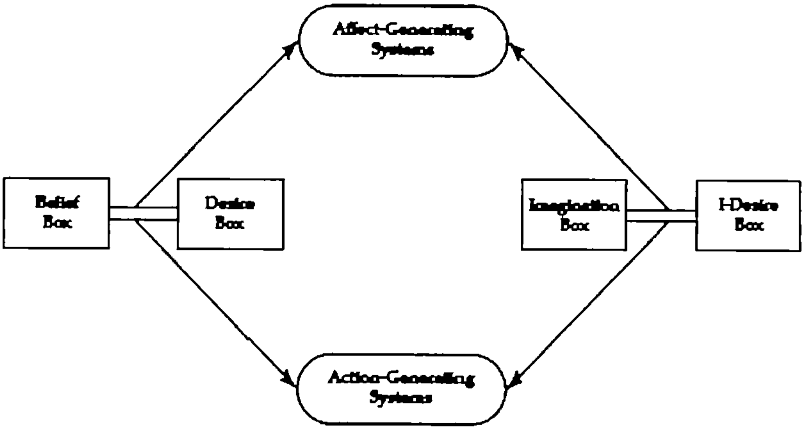
What I call the *replacement account* does hold that imagining has an unmediated role in generating pretend play. This account modifies the standard belief-desire explanation of action in a different way from the conditional belief account, which made imagining a *source* of (conditional) beliefs. On the replacement account, imaginings take two forms—one form corresponding to beliefs and one to desires—and they effectively replace beliefs and desires in producing action. Serious action, then, is the action motivated by beliefs and desires; pretense is motivated by *replacement beliefs* and *replacement desires*. To talk about these theoretical posits concisely, I extend the Doggett and Egan terminology by calling imaginative replacement beliefs "i-beliefs" and replacement desires "i-desires." (Doggett and Egan only write of "i-desires" and simply use "imagination" for the posit I call "i-belief."[14])

---

[13] I use "cognitive attitude" and "conative attitude" in their usual senses. A cognitive attitude is an attitude that represents the world or some portion of it as *being* a certain way; beliefs, hypotheses, assumptions for the sake of argument, and the like are all cognitive attitudes. Conative attitudes, such as desires and wishes, represent the world as *to be made* a certain way. (This formulation borrows from Nishi Shah and Velleman, "Doxastic Deliberation," *Philosophical Review*, cxiv, 4 (October 2005): 495–519.)

[14] Amy Kind independently hit upon using "i-belief" in discussing this family of views.

Doggett's and Egan's diagram clarifies the replacement account's structure.[15]



The Imagination Box (i-beliefs) and I-Desire Box do what the Belief and Desire Boxes do, except when the former boxes do it it counts as pretense, whereas when the latter boxes do it it counts as serious action. If Sally asks Grammy for ice cream while wanting ice cream and believing she can get it by asking Grammy, that is serious action. If Sally says a "magic" spell, i-desiring to fly to a magic kingdom and i-believing she can get there through the spell, that is pretense. (This characterization raises the question of what the *difference* is supposed to be between imagining and belief, a question which I think no replacement theorist answers satisfactorily. But since that is not what is at issue in this paper, I will not pursue that criticism here.[16])

In addition to Doggett's and Egan's paper, the replacement account can be found in Velleman and in Currie and Ravenscroft. What are their motivations?

Velleman is motivated to understand what distinguishes belief from imagining, and finds by considering examples that this is a difficult problem. The reason is that "imaginings" seem to him to motivate behavior in relation to "wishes" in the same way that beliefs motivate in relation to desires. To give one of Velleman's examples, imagining she is an elephant and wishing to take a drink, a child is motivated to pretend to drink from a "pail" (where a chair is the prop for the pail) with her "trunk" (arm). Thus, for Velleman, examples show that imaginings (i-beliefs) and wishes (i-desires) can replace beliefs and desires in motivating action.

---

[15] Doggett and Egan, *op. cit.*, p. 4.

[16] I take up that issue in "The Motivational Role of Belief," *Philosophical Papers*, xxxviii, 2 (July 2009): 219–46.

Currie and Ravenscroft use the terms "belief-like imagining" and "desire-like imagining." Their motivation is simulation theory, which is a theory about how people understand decisions and predict the behavior of others. In short, simulation theory says we understand the minds of others by running our own belief, desire, and other mental machinery "offline." So for Currie and Ravenscroft, the theoretical posits of belief-like imagining and desire-like imagining originated as concepts of "offline" states that people use in understanding the minds of others. Their thought about pretense seems to be that, since they are positing such states already, it would be nice to use them for something else as well, such as explaining pretense.[17]

Doggett and Egan are motivated by two things. First, they see the Nichols and Stich architecture as insufficient (for many of the same reasons that I do). Second, they think positing i-desires is parsimonious insofar as they can do double duty: explain pretend behavior and explain the emotions people have in response to imagining, especially in relation to fictional drama. If desires can explain emotion, so can i-desires. On this second point, Currie and Ravenscroft and Velleman seem to be in agreement.

There are, in sum, three points common to the replacement accounts of pretense:

(1) i-beliefs and i-desires are propositional attitudes.
(2) i-beliefs and i-desires influence action generation in the manner of beliefs and desires.
(3) i-beliefs and i-desires influence emotion in the manner of beliefs and desires.

The motivations for the replacement view amount to examples, a preexisting theory (simulation theory), and apparent parsimony.

This account seems even more intuitively compelling than the conditional belief account. But it fails just as badly. First, as with the conditional belief account, the replacement account cannot explain *layering*. Second, due to the binary nature of the architecture posited, the replacement account cannot account for the continuity we see in forms of pretense; that is, it cannot account for *semi-pretense*. Third, the theoretical commitments that replacement theorists put on their posits, i-beliefs and i-desires, conflict in an important class of cases. But, as before, understanding how this account fails leads to constructive suggestions about how to structure a better theory of imagining and pretense.

---

[17] To be fair, Currie and Ravenscroft discuss four different ways pretense could be motivated. All of them, however, are modifications of desire-belief psychology, so they do not incorporate the points about pretense I develop in section III.

Criticism 1 (layering): Let us return to "Rebel Without a Cause," in which Dean's and Wood's high school student characters pretend to be adults who consider buying a house. The replacement account posits in Dean an i-desire to escape (as Jim) with Judy and an i-belief that by entering a deserted mansion he can do this. That explains the first layer of pretending to be Jim in the mansion. But what i-belief/i-desire pair causes the behavior of saying, "Should we rent or are we in a buying mood, dear?" Dean performs that action in the character of an adult, not as high school Jim. The problem with the replacement account is that it has no resource to explain the second-layer pretense actions, since the i-beliefs and i-desires are dedicated to the first-layer actions. Furthermore, the i-beliefs and i-desires cannot represent both layers, since that would result in incoherent i-beliefs, which will not explain anything.

We can put this criticism as a dilemma. Either (i) i-beliefs and i-desires explain the first "Jim" layer of pretense, or (ii) they explain the second "adult" layer. If (i), then the second layer is mysterious. If (ii), then this is no longer a case of pretense layering, but a case of pretense changing; under (ii), there will no longer be any sense in which Dean *keeps* pretending to be the high school student while pretending to be the adult.[18] There is no layering.

Criticism 2 (binarity versus semi-pretense): If beliefs and desires explain plain action, and if i-beliefs and i-desires explain pretense, what explains semi-pretense? As noted, there is a continuum between full pretense and plain action. But the replacement account is binary; there are only two pairs of boxes, which suggests that the plain action/pretense distinction should also be binary. But it is not. Jennifer holds up ten fingers in response to the swan dive. Do we say that she desired to indicate her approval and believed that holding up ten fingers would indicate approval? Or do we say that she i-desired (*qua* judge) to give the highest score and i-believed that ten was the highest score? The first option leaves the similarity of Jennifer's action to judging behavior unexplained. The second leaves unexplained the difference between what she does and what she would do in *full* pretense, where she pretends to hold up a placard and speaks in more officious tones.

---

[18] The replacement theorists could posit i-i-beliefs and i-i-desires, but this would only add another pair of boxes that looks just like the other pairs, with only the theorist's labels to distinguish them; there would be no sense in which the idea of *layering* is preserved.

The replacement account has to treat plain action and pretense as binary. Its only resort, then, is to say that semi-pretense is not a truly distinct phenomenon. It is an illusion that stems from the fact that plain action can be playful and pretense can be done more or less seriously. Whether that claim is true depends ultimately on the underlying architecture of imagination. Is it binary or not, and if so, in what ways? But even if the response is true—I think it is not—the replacement account leaves *manner* unexplained. Why is some plain or "serious" action playful, having pretend-like qualities? Does imagination have nothing to do with this?

Criticism 3 (conflicting theoretical demands): Doggett and Egan and Currie and Ravenscroft explicitly state that i-desires explain emotional reaction to what is imagined (i-believed), where what is imagined can have fiction, such as novels and films, as its source. Why do I feel angry at Iago? The replacement theorist says it is because I i-believe his deceit will cause catastrophe and i-desire Desdemona and Othello to have a happy life. The general principle seems to be that i-desiring that p yields negative emotions in conjunction with i-beliefs that $\sim p$ (or that $\sim p$ is likely) and positive emotions in conjunction with i-beliefs that p (or that p is likely).

The problem of conflicting theoretical demands comes when people pretend to do things that run *contrary* to the emotions that a given imagining engenders. Limburger cheese, for example, may be so repulsive to me that *imagining* eating it engenders disgust. Nevertheless, I can pretend to eat an imaginary plate of Limburger, even as I feel disgust about what I am pretending. For such a case, the replacement theorist has to posit an i-desire *to eat the Limburger* (to explain the pretense behavior) and an i-desire *not to eat Limburger ever* (to explain the disgust). The general problem is this: whenever someone pretends an action contrary to her emotional inclinations about what she is pretending, the replacement account has to posit conflicting i-desires—an i-desire to explain the pretense and a conflicting i-desire to explain the emotion. This raises a puzzling question. Why doesn't the i-desire *to* eat Limburger generate the emotion and the i-desire never to eat Limburger generate the pretense action? In that case, I would feel excitement about the imaginary Limburger in front of me but pretend *not* to eat it, which is the opposite of what I did. More generally, when there are two conflicting i-desires, which one gets to generate the pretense action, and which one gets to generate the emotion? The replacement account as it stands has no answer to this question, so it needs elaboration. Thus, replacement theorists cannot claim parsimony in positing i-desires as quickly as they thought. Worse yet, the replacement account says that i-desires generate pretense action and emotion, but we have seen that for some cases it will have to say (without explanation so far) that sometimes they do not generate action and sometimes they do not generate emotion. That seems to conflict.[19]

* * *

What constructive suggestions can we glean from these criticisms? The replacement account falters because it *separates* the veridical boxes sharply from the nonveridical boxes. We can remedy this by saying that imaginative representations can *integrate* with veridical representations. If integration were to happen at the higher cognitive level, the level of belief, the subject would become confused, like a schizophrenic. But it may be that integrating veridical and imaginary representations is possible at the level of perceptual and motor representation. We know from Stephen Kosslyn's long research program—not to mention just doing phenomenology—that there are indeed perceptually formatted imaginary representations, or mental images. The integration of those imaginings with veridical perceptions—allowing imagery seamlessly to take the roles of percepts—may alleviate the problems caused by the binary replacement picture. This view would require moving the imaginings most immediate to the causation of pretense away from the category of propositional attitude and into the category of perceptual imagery (and, I argue, motor imagery). But that is for the best, if we are to explain the *emotions* generated in response to imagining, which are a further problem for the replacement account in any case. There is reason to think that imagining in, say, *visual* or *auditory* detail has far greater emotional impact than bare, propositional imagining. But that does not mean we should write propositional attitudes out of the story. Rather, as suggested, cognitive and conative (propositional) attitudes help cause perceptually formatted (or other spatially rich) imaginings, but *those* imaginings are most immediate to the production of both pretense action and emotion. At the same time, cognitive and conative attitudes that comment on the rich imaginings keep track of which *layer* of pretense a rich imagining represents. And, to preview further, semi-pretense could be what happens when spatially rich imaginings influence behavior *without* the sustaining, reflective involvement from higher cognitive and conative attitudes.

I turn now to clarifying what forms of *integration* are relevant to these claims; that is, I now give my positive theory.

---

[19] For similar criticisms, see Eric Funkhouser and Shannon Spaulding, "Imagination and Other Scripts," *Philosophical Studies*, CXLIII, 3 (March 2009): 291–314.

III. THE INTEGRATED IMAGINATION THEORY

Let us turn the constructive suggestions into formal claims and apply them to reveal how imagination is involved in producing action.

Claim 1: Imagining influences pretense behavior without the mediation of belief.

Claim 2: Cognitive and conative attitudes, like beliefs and desires, help cause what one imagines.

Claim 3: Cognitive and conative attitudes comment on and constrain imaginings and how they influence pretense.

Claim 4: The imaginings that are most immediate to the production of pretense are spatially rich; they are perceptually formatted or structured as representations of bodily movement.

Claim 5: Perceptually formatted imaginings can be integrated with veridical perception, especially visual or auditory perception.[20]

Claim 6: Imagining arouses emotions largely in the same way that perception does, as imaginings play out on perceptual cortices that feed into the brain's emotional centers.[21]

Let me clarify some points.

Claim 1 does not entail that beliefs have no role in causing imaginings or in generating pretense; rather, when it comes to generating pretense actions (and other actions), imaginings of the type I posit are closer to the generation of action insofar as they are causally downstream from beliefs. That is the point Claim 2 emphasizes.

Claim 3 attributes more roles to beliefs and desires (or other attitudes) in the production of pretense than simply causing imaginings. Beliefs and desires can be *about* what one imagines and can comment on the *value* to the agent of a particular imagined action. This is important to explaining how ~~choice among~~ multiple imagined possibilities happens. One can imagine doing an action and then believe that doing *that* would be good, where the imagining gives content to the belief.[22] The comments on imaginings by beliefs and desires about the value of an imagined action then constrain what further actions and events one imagines.

---

[20] Most of my examples are visual, but the claim holds for audition, like imagining a voice in the wind.

[21] Tamar Gendler discusses emotional response to imagining under the heading "contagion" in "Imaginary Contagion," *Metaphilosophy*, xxxvii, 2 (April 2006): 1–21. The term is "contagion" because imagining seems to infiltrate the role normally occupied by belief. But I object to the term "contagion" because I think emotive responses to imagining are a feature, not a bug. They may be problematic sometimes, but they do more good than harm, since they give orientation with respect to the imagined future.

[22] John Perry gives an interpretation of David Kaplan according to which demonstrative constituents of representations (for example, *that*) have their contents partly determined by a "directing intention," which in turn connects to the world in many cases via perception. See Perry, "Directing Intentions," in Joseph Almog and Paolo Leonardi, eds., *The Philosophy of David Kaplan* (New York: Oxford, 2009), pp. 187–201.

A helpful analogy comes from the history of geometry. Prior to Descartes' invention of the Cartesian plane, geometric proofs required diagrams. Yet proofs were accompanied by comments in the margins that indicated which properties of the diagram were relevant to the proof. The imagining I posit in this paper is analogous to the spatial diagrams—and the drawing of them—while beliefs, desires, and propositional imaginings are analogous to the comments. Furthermore, consider what went on in the geometer's *mind* when reading such proofs. There was spatial perception of the diagram, but there were also beliefs about the diagram, caused by the marginal comments. But the beliefs had their contents tied to the diagram *through* relating to the percepts; there is no other way for beliefs about the diagram to be had. Now consider the mind of the geometer who, with eyes closed, *remembers* a proof from before and knows what theorem it proves. This is a clear case of perceptual imagining with higher beliefs commenting on the perceptual imagining. I take this example to be an independent existence proof of the sort of mental structures to which I appeal in explaining pretense: spatially rich imagining with comment and constraint by higher attitudes.

Claim 4 appeals to the traditional idea of mental imagery but also emphasizes that anticipatory motor plans, or *forward models*, can be imaginative.[23] That means one can construct an action in one's mind as an imagining without committing to its ever happening; this is a motor plan in potential. A basketball player who imagines her free-throw motion from a first-person perspective prior to taking the shot is imaginatively constructing such a forward model.[24]

Claim 5 is the most radical departure from the two theories just considered, which had imagining separated off in distinct boxes.[25] Claim 5 allows for mental images to be integrated with the perceptual field. Try an experiment. Is there an empty chair near you? If so, look

---

Perry's focus is language, but the point can be extended to mental representation. I say here that a visual (or other sensori-form) imagining can take the place of perception in giving content to a demonstrative constituent of belief or propositional imagining.

[23] This idea of motor imagining is present in Currie and Ravenscroft, *op. cit.*, as is mental imagery. But they do not incorporate these notions into what they say about the production of pretense.

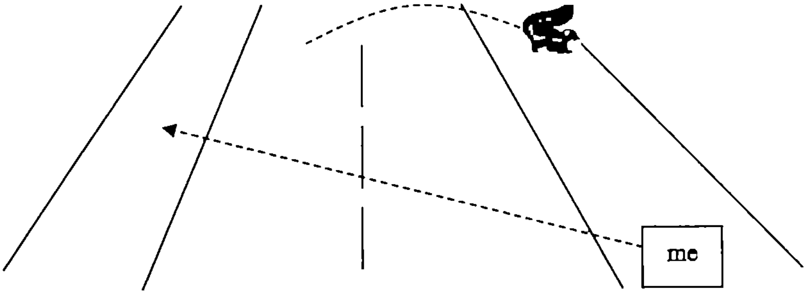
[24] Thanks to Juli Huddleston for this example. She also gave me the example of the pole-vaulter who visualizes her vault from start to finish from the first-person perspective before competing.

[25] Boxologists may respond that the "boxes" are just metaphors. But the metaphor is misleading

at it, and imagine Santa Claus sitting in *that* chair. Is your mental image of Santa Claus in a separate box from your other perceptions? Or is it integrated with the perceptual manifold before you? If you think the latter, you tend toward my view. The projected geometry of the imaginary Santa Claus is in the same represented space as the perceived chair and is represented as bearing rich spatial relations to it—much richer than the simple <*on the chair*> of which propositional imagining is capable. Such integrated imaginings are central to the production of pretense.[26]

Claim 6 is fairly clear: whatever the pathways are by which perception excites emotion, those are the pathways by which imagining excites emotion. But again, that does not mean that higher cognitive and conative attitudes are otiose. If you visually imagine a cousin of yours, part of what makes it possible for this to be an imagining *about your cousin* are beliefs you have about your cousin: what he looks like, what his manners are, and so on. Thus, insofar as the identity of individuals in your imaginings makes a difference to how those imaginings make you feel, higher cognitive attitudes make a difference as well. Still, the road from belief to emotion goes through imagination—at least in imaginative fiction and pretense.[27]

Now I turn to an illustration.



I was jogging down the street in Medford one day, when I suddenly saw a skunk in front of me. Naturally, I did not want to get sprayed. So I visualized the skunk spraying, imagined myself running across the street to a distance beyond where I imagined the skunk spray going, and then ran across along the route I had imagined. I felt a sense of relief once I had passed, having felt nervous before.

---

[26] Of course, not all images are so integrated, since one can close one's eyes and imagine. But many are.

[27] My view here pushes against Kendall Walton's view of fear. He holds that genuine fear requires *belief*. But see Robert Zajonc, "On the Primacy of Affect," *American Psychologist*, xxxix, 2 (1984). 117–23. Also, consider this: someone watches a Bigfoot film, not believing any of it, and then is scared later by a friend in a Bigfoot suit who jumps out from behind a tree. Is that not fear? It is possible (and I think actual) that fear formation is faster than belief formation. Walton's view excludes this possibility.

The skunk representation was a veridical perception that made me feel nervous in the way that perceptions of animals can. The skunk-spray representation, however, was an imagining. Crucially, this imagining was visual and was integrated into my visual field (Claims 4 and 5). Furthermore, I constructed my motor imagining—the path that I would take—*in relation to* the visually imagined spray; I used the imagined visualization of the spray to choose a path across the perceived street. (Had the visualization not been integrated into the perceptual manifold, I would not have known how much of the street to cross.) Of course, I would not have imagined the skunk spraying in the first place if I did not *believe* that skunks spray (Claims 2 and 3), so belief in fact was involved in the imagining and ultimately in the action, but it was causally upstream from the imagining in relation to the action (Claim 1).

So far this is an example of how imagining is involved in plain action. What about pretending? Suppose I had been jogging along and decided to *pretend* there was a skunk in my path, even though I had not seen one. My bodily motions would have been the same; I just would have been pretending to avoid a skunk instead of seriously avoiding a skunk.

The crucial point from the cases is this: *the structure of representations that enables me to pretend to avoid the spray of an imaginary skunk is almost entirely the same as the structure of representations that enables me seriously to avoid the potential spray of a real skunk.* That being said, there are two mental differences between the pretense and serious skunk avoidances. First, the veridical perception of the skunk comes with anticipation of further recalcitrant experience, or experience beyond voluntary control, such as catching my foot if I ran into what the percept is of; the imaginary skunk image does not involve anticipation of recalcitrance.[28]

---

[28] This point could be expanded greatly. The question is: what constitutes the *valence* that differentiates imagery from perception, given that the representational format is the same? In an earlier draft, I said that perception was more "solid" and that it "obscured" what was behind it. But Dagfinn Føllesdal pointed out (personal communication) that it is in principle possible to *imagine* something that obscures what is behind it. So *lacking solidity* and *not obscuring* are contingent features of the imaginary image. "Anticipation of further recalcitrant experience" gets closer to defining the valence of veridical perception (see Føllesdal, "Existence, Inexpressibility, and Philosophical Knowledge," *Grazer Philosophische Studien*, LXXIV, 1 (2007): 273–90, who is interpreting Husserl). But I argue in "What Is a Judgment?" that we need to identify *asymmetric* relations between attitude valences to define them; on my view, veridical representations *govern* and *ground* representations with other valences.

Second, the higher cognitive attitudes that comment on the perceptual manifold include, in the serious case, a belief that *that is a skunk*, where the <*that*> constituent has its content given by the skunk percept. The higher cognitive attitudes include, in the pretense case, a belief that *that is an imaginary skunk*, where the <*that*> constituent this time has its content given by the skunk mental image. (Alternately, we may say that a propositional imagining comments on the visualization of the skunk, where "propositional imagining" refers to what the other theorists were attempting to talk about with their imagination boxes. Although I reject the other theories, I do believe there are propositional imaginings; a good example would be the long-standing imaginative attitude that J. R. R. Tolkein had that *elves are immortal*, which guided how he wrote many of the episodes in *Lord of the Rings*. The propositional imagining in my pretense would then have the content *that is a skunk*.)

Claims 1 through 5 apply to the pretense case as to the serious case. We may add that, insofar as I feel emotional excitement in the pretense, it is because of connections that visually formatted representations have to the brain's emotional centers, as maintained in Claim 6.

Let us explicate the main thesis of this paper in relation to these six claims. The Active Imagination Thesis posits a type of imagining that is a forward model of action in the world. The imagination constructs this model in relation to objects of perception and in doing so seeks to satisfy the goals set by higher attitudes, such as beliefs and desires. In pretend play, two shifts can happen. First, some of the objects in the perceptual field, in relation to which motor commands are constructed, can be imaginary mental images. The sword I duck in make-believe is the one I visualize in your (actually empty) hand as you make a slicing motion aimed at my head. Second, the higher attitudes can comment on the actor/subject, whose motions the forward model represents, in a way that makes that actor/subject out to be someone or something else. If I am pretending to be a prince who ducks the sword, I have a higher imaginative attitude commenting on the subject of the ducking with the content *I [the subject] am a prince*. In some pretense, one of these shifts happens without the other, as when I pretend *in my own character* to give a speech to an audience in a room of actually empty chairs (visualization of new entities without change of acting subject), or when one is at a fancy high-table lunch and imagines oneself to be an actor in a Gilbert and Sullivan musical[29] (change of subject without additional visualizations).

---

[29] Thanks to Daniel Dennett for this example.

Let me add one elaboration to extend the theory to pretense involving *imitation*.

Suppose you see a bird flying through the air and decide to imitate it. You then produce a flapping motion with your arms to mirror the perceived motion of the bird's wings and a chirp of your voice to mirror the bird's heard chirp. This is imitation, or *mimesis*, and humans are uniquely good at it.[30] In pretense, what is imitated need not be directly before one in the external world. One can imitate a mental image or other imagined content. What happens, for example, in a game of charades? You read a card that says "flying bird." A mental image of a bird flapping its wings pops into your mind. So, playing the game, you imitate the motion of the bird that is represented in your visual imagination by flapping your arms. If it pops into your auditory imagination that the bird chirps, you chirp. In sum, the visual and auditory mental images in the charades play the role that the perceptions did when you *saw* the bird; they give you something to imitate. It should be no surprise that there is a parallelism, since imagery plays out on the same cortices as perception. What is needed to motivate such pretense is the desire to act out—copy the motions of—whatever is going on in imagination. Let us call this *mimesis-based pretense*.

The way mimesis-based pretense meshes with the Active Imagination Thesis is as follows: in addition to providing objects in relation to which forward-model imagining constructs actions, mental imagery can set the structure of motion that the forward model models. There are thus two different roles for mental imagery in action and pretense; let us call the first the *object-acted-upon* role and the second the *object-imitated* role. Our two paradigm examples of these roles are the swinging sword and the flying bird. The visualized swinging sword gives, via perceptual imagining, a spatially represented object for the forward model to make my body avoid; the visualized flying bird gives a motion for the forward model to imitate.

The Evidence for Neuropsychological Plausibility

My account is built out of a number of components, many of which are supported by research programs in cognitive neuroscience. What follows is a small sample of important results.

---

[30] The capacity for (precise) imitation is distinctive of our species. See Joseph Henrich's and Richard McElreath's "The Evolution of Cultural Evolution," *Evolutionary Anthropology*, XII, 3 (2003): 123–35, and Michael Tomasello's *The Cultural Origins of Human Cognition* (Cambridge: Harvard, 1999), but also see chapter four of Aristotle's *Poetics*

Fact 1: Mental imagery exists.
Fact 2: Mental imagery is processed largely in the same regions of the brain as perception.
Fact 3: Forward models are implicated in the planning and performance of motor actions.
Fact 4: The brain's perceptual regions have a fairly direct role in activating emotional centers and processes.

Kosslyn's research program establishes Fact 1.[31] The way to interpret Fact 1 is to say that imaginatively generated mental representations exist that (1) are structured like perception and (2) depict spatial properties in analogue fashion. This is not a "pictures in the mind" view; rather, *however* the brain encodes percepts, it can encode imaginary representations in like manner.

The evidence for Fact 2 is intertwined with the evidence for Fact 1. Kosslyn et al. show how visual area 17 is activated both in perception and in imaginative imagery. The same sorts of contents, whether perceived or imagined, yield similar activation patterns. Other results point in the same direction. Kathleen O'Craven and Nancy Kanwisher show how the fusiform face area (FFA) is activated both by seeing faces and by imagining them. Jonathan Winawer, Alexander Huk, and Lera Boroditsky showed subjects pictures that had *implied* motion (for example, a runner in mid-stride) and found that these images activate the same direction-selective circuits that process real motion; I take this to show that *imagining* motion on seeing the photograph is processed similarly to *seeing* motion. These few studies are the tip of an iceberg supporting Fact 2.[32]

Fact 3 follows primarily from studies that assume the existence of a forward model in the central nervous system and use that assumption to derive accurate predictions about motion. For example, the assumption that the brain uses forward models to plan action can yield accurate estimates about the extent to which a subject will over- or underestimate an arm motion toward an object, when resistance or help is applied to the arm.[33]

---

[31] Summarized in Kosslyn, William Thompson, and Giorgio Ganis, *The Case for Mental Imagery* (New York: Oxford, 2006).

[32] References: Kosslyn et al., "The Role of Area 17 in Visual Imagery: Convergent Evidence from PET and rTMS," *Science*, CCLXXXIV, 5411 (April 1999): 167–70; J. Winawer, A. Huk, and L. Boroditsky, "A Motion Aftereffect from Still Photographs Depicting Motion," *Psychological Science*, XIX, 3 (March 2008). 276–83; K. O'Craven and N Kanwisher, "Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions," *Journal of Cognitive Neuroscience*, XII, 6 (November 2000): 1013–23.

[33] See D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An Internal Model of Sensorimotor Integration," *Science*, CCLXIX, 5232 (September 1995): 1880–82; Wolpert,

Fact 4 also has extensive evidence. A representative research program is that of Joseph LeDoux, who finds there is a "low" and a "high" road from sensory cortices to the amygdalae, the brain's fear centers. The low road is fast and sloppy and is responsible for fear responses to things that only loosely *resemble* dangerous objects; LeDoux posits sub-doxastic appraisals along the low road—representations of particular dangerous types, like snakes—that trigger fear responses. The "high" road is slower but has more precision and nuance. The high road can extinguish fear responses to stimuli that are not actually dangerous. In a way, Fact 4 should be obvious prior to any evidence from neuroscience; we have always known that sights and sounds stimulate emotion. But we are becoming much clearer on *how* that happens.[34]

Facts 1, 2, and 4 yield my view of imagining and emotion: the perceptual aspect of imagining is central to emotional stimulation, activating emotional responses by the same pathways as perception itself.[35] We need no i-desires to explain the emotions that fiction incites.

Finally, there are two empirical claims to which my account commits me that, as far as I know, have not been rigorously tested. First, I hold that, in pretense, perceptual imagining is integrated with perception. That may seem obvious to phenomenological reflection, but we should demand corroboration. This claim, however, will be hard to test by the usual means; much evidence for the brain activation of imagery comes from fMRI studies, but the sort of pretend play I discuss, which incorporates both perceived and imagined objects integrated in the same manifold, can hardly be done *well* in the scanner.[36] Second, I say that a mental image in the object-imitated role can

---

R. C. Miall, and M. Kawato, "Internal Models in the Cerebellum," *Trends in Cognitive Sciences*, II, 9 (September 1998). 338–47; and (for a broad extension of the research program) R. Grush, "The Emulation Theory of Representation: Motor Control, Imagery, and Perception," *Behavioral and Brain Sciences*, XXVII, 3 (June 2004): 377–442.

[34] LeDoux summarizes this research in *The Emotional Brain* (New York: Simon & Schuster, 1996)

[35] This is essentially the view that Timothy Schroeder and Carl Matheson advocate in "Imagination and Emotion," in Nichols, ed., *The Architecture of Imagination* (New York: Oxford, 2006), pp. 19–40.

[36] One of the very first studies on mental imagery bears on this integration claim. C. W. Perky had subjects look at a screen with instructions telling them to imagine, say, a banana; then, unbeknownst to the subjects, a banana would be projected faintly on the screen. Perky found that subjects took the real banana image for a mental image. See "An Experimental Study of Imagination," *American Journal of Psychology*, XXI, 3 (July 1910): 422–52. This result, if legitimate, supports my view that mental images can integrate with the visual field; otherwise subjects would never confuse a percept with an image.

have influence on the construction of the forward model of bodily motion. Research into mirror neurons may support this view in some way, but as far as I can tell, it is still too early to say.[57] Nevertheless, we know that imitation of perceived actions happens somehow; my view is that pretense proceeds by replicating much the same process, whatever it is, with the difference being that a mental image of the object imitated substitutes for a perception.

Application to Semi-Pretense and Layering

As she sits next to Leroy at the pool, the mental image of a judge pops into Jennifer's mind, and this image takes the *object-imitated* role in the construction of action. Once the image has taken this role, it is implicated in forming the forward model of how she might behave. If that behavior is supported by her other attitudes—say, a desire to joke around with Leroy—she will act out the judge-imitating forward model, delivering verdicts like "Ten!"

Now, there can be various causes of the judge image that occurs to her. If an explicit desire or intention to *pretend* causes the image, the process that follows will be full pretense. Many actions occur in imagery, however, with *other* causal sources (besides desire or intention) and take the object-imitated role nonetheless. That is the crucial point. Images can occur due to habit, obsession, resemblance of a current situation to a past one, misunderstanding, and so on. When the source of the mental image in the object-imitated role is something other than a desire or intention to pretend, what results may be semi-pretense, in which the image still affects the forward model of action in subtle (unintended) ways. Thus, semi-pretense can be defined as follows:

> Action *A* by subject *S* is *semi-pretense* if, and only if:
> (i) *S* has a mental image *M* in the object-imitated role that influences the action recipe[58] or forward model that eventuates in *A*, and
> (ii) *S* lacks the desire or intention to pretend to do the action represented in *M*.

On this view, semi-pretense is what happens when the images that swim in our minds infect our actions without our explicitly intending them to do so. Such actions may even become habit. If they do, there may eventually cease to be any sense in calling them "pretense," since the mental images causing them have long since vanished. In that case, the forward model of action lives on through habit, despite the death of the image that gave it birth. The ghost of the object imitated pervades the bodily motions of the habitual actor long into the future, without any conscious knowledge on his part.

---

[57] For the case for pervasive involvement of mirror neurons, see Alvin Goldman's *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading* (New York: Oxford, 2006).

[58] By "action recipe" I just mean a structure of behaviors that add up to a larger action, for example, <push ON, push OPEN, open DVD case, insert DVD>. Action recipes are a level more abstract than forward models, since forward models code specific bodily motions, while, for example, the action recipe just mentioned could be executed whether the ON button is on the right or left, that is, with different bodily motions.

Let us turn to pretense layering. I have emphasized throughout that beliefs comment on the contents of perceptual imagining and motor imagining. Whenever a given action is constructed in spatially rich perceptual and motor imagining, beliefs (or propositional imaginings) comment on the story to which *that* pretense action belongs. To return to our analogy with geometry, marginal comments can track whether a given diagram belongs to the main proof or sub-proof. Thus, when James Dean says, "Should we rent or are we in a buying mood, dear?" the most immediate cause of this action is a forward-model imagining that is influenced by a mental image of an adult in the object-imitated role; what makes that action second-layer pretense, as opposed to first-layer, is the set of beliefs and higher imaginings that places the image and action in a larger story. Those attitudes then further cause and guide what spatially rich imaginings come next in Dean's mind, leading him up the stairs with Judy.

IV. CONCLUSION: A STRUCTURE TO GUIDE US

Imagination is, ultimately, the production of a structure that *can* guide action. That action must occur in the world we take in via a flood of information through the senses. Thus, there must be forms of imagining that integrate seamlessly with our perceptual representations, if imagination is to be of any use at all. That is why, I believe, imagining plays out on the sensory cortices of the brain; that is also why propositional imagining is not *sufficient* to explain the variation and complexity of pretense behavior—and other behavior. In the Nichols and Stich theory, there are three boxes between imaginings and the "Action Control systems" and two boxes between imaginings and perceptual processes, which (to judge from the arrows) do not interact with imagining much at all. My theory could not be more different. The forms of imagining I posit integrate with perception and guide action.

But let me emphasize once again: this does not mean that beliefs and desires have nothing to do with *choice* of either pretense or serious action. Rather the beliefs, desires, and other attitudes that comment on and constrain imaginative representations are crucial to determining whether and when imagined actions are initiated. But the spatial resolution of imagination makes it suitable to producing structures that can dictate the geometric unfolding of bodily motion in the world. Imagination is where the action is.

NEIL VAN LEEUWEN

University of Johannesburg and
Georgia State University

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA
ALIPORE CAMPUS

SELF-CONSCIOUSNESS AND IMMUNITY*

One of the most seminal contributions to the understanding of self-consciousness over the last half century has been Sydney Shoemaker's articulation of the idea that we are "immune to error through misidentification relative to the first-person pronouns" (IEM).[1] Along with related ideas developed by Wittgenstein, Castenada, Evans, Perry, and Pryor,[2] IEM has proven to be extremely fertile in stimulating insights into the first-person perspective, "the distinctive way mental states present themselves to the subjects whose states they are."[3] Moreover, Shoemaker's formulation of the idea has motivated significant interdisciplinary research into self-consciousness.[4]

Since first formulating his position, Shoemaker has done much to elaborate upon IEM and related notions. For more than four decades he has been perspicaciously developing his ideas on identification-freedom, introspection, self-knowledge, and the self-intimation of mental states. Although some aspects of Shoemaker's views on immunity have been disputed, IEM itself has never been severely threatened by any empirical challenge.[5]

---

*We are extremely grateful to Frédérique de Vignemont, Tim Bayne, Thomas Metzinger, Olaf Blanke, Shaun Gallagher, David Rosenthal, Valeria Petkova, several participants in the ASSC 13 Conference (in Berlin), and to this JOURNAL's anonymous referees for their helpful comments on previous versions of this manuscript. Work on this manuscript was, in part, funded by National Science Council of Taiwan research grants 97-2410-H-004-154-MY3 and 97-2410-H-002-184-MY3.

[1] Sydney Shoemaker, "Self-Reference and Self-Awareness," this JOURNAL, LXV, 19 (October 3, 1968): 555–67. In previous work he had attempted to make a similar point by referring to certain self-ascriptions that are "noncriterial." See Shoemaker, *Identity, Cause, and Mind* (New York: Oxford, 2003), p. 2.

[2] See Ludwig Wittgenstein, *The Blue and Brown Books* (New York: Oxford, 1958). Also see Gareth Evans, *The Varieties of Reference* (New York: Oxford, 1982); James Pryor, "Immunity to Error through Misidentification," *Philosophical Topics*, XXVI, 1–2 (1999): 271–304; and additional representative works in Andrew Brook and Richard DeVidi, eds., *Self-Reference and Self-Awareness* (Philadelphia: John Benjamins, 2001).

[3] Shoemaker, *The First-Person Perspective and Other Essays* (Cambridge: MIT, 1996).

[4] See Jose Bermudez et al., eds., *The Body and the Self* (Cambridge. MIT, 1995); Bermudez, *The Paradox of Self-Consciousness* (Cambridge: MIT, 1998); and Masaharu Mizumoto and Masato Ishikawa, "Immunity to Error through Misidentification and the Bodily Illusion Experiment," *Journal of Consciousness Studies*, XII, 7 (2005): 3–19.

[5] For example: several essays in Brook and DeVidi, eds., *op. cit.*, Colin McGinn, *The Subjective View* (New York: Oxford, 1983), pp. 45–55; Evans, *op. cit.*, pp. 188–91; and Bermudez, *op. cit.*, pp. 6–8.

0022-362X/11/0802/078–099

Perhaps the most substantial empirical challenge thus far attempted has been Campbell's[6] claim that schizophrenic thought insertions, understood in terms of the Frith[7] monitoring model, might serve as a counterexample to IEM. Gallagher and Coliva have defended IEM by (among other things) arguing that since schizophrenic thoughts are still within a patient's stream of consciousness, IEM is not violated.[8] They hold that, as a matter of conceptual truth, "if a subject is introspectively aware of a certain mental state, then she herself is having it and, therefore, that mental state is her own."[9]

In this paper we argue that IEM fails. In section I, we adumbrate Shoemaker's version of IEM along with related concepts central to his understanding of self-consciousness. We also reject the interpretation of IEM as a tautology, and propose to treat it as a hypothesis. In section II, we present a clinical case—somatoparaphrenia—and in section III we describe an experiment with healthy subjects—the Body Swap Illusion. In the former case, patients represent experienced sensations as belonging to someone other than self. In the latter, an illusion is created whereby subjects feel that they can shake hands with themselves. The one concerns bodily sensations; the other, the sense of agency.[10] Both cases reveal that IEM lacks modal force: what IEM says cannot happen, can happen. In section IV we respond to possible criticisms of our position. In a concluding section we emphasize that in order to account for the phenomena which seem to defy IEM-based expectations, there is a need to distinguish the ownership of mental states from the ownership of body parts. Moreover, concerning the former, there is a compelling need to distinguish between mental states that are instantiated and mental states that are *represented as belonging to oneself*.

---

[6] John Campbell, "Immunity to Error through Misidentification and the Meaning of a Referring Term," *Philosophical Topics*, xxvi, 1–2 (1999): 89–104.

[7] Christopher Frith, *The Cognitive Neuropsychology of Schizophrenia* (Hove, UK: Erlbaum, 1992).

[8] Shaun Gallagher, "Self-Reference and Schizophrenia: A Cognitive Model of Immunity to Error through Misidentification," in Dan Zahavi, ed., *Exploring the Self: Philosophical and Psychological Perspectives on Self-Experience* (Philadelphia: John Benjamins, 2000), pp. 203–39; Annalisa Coliva, "Thought Insertion and Immunity to Error through Misidentification," *Philosophy, Psychiatry and Psychology*, ix, 1 (2000): 27–34; and Coliva, "On What There Really Is to Our Notion of a Thought," *Philosophy, Psychiatry and Psychology*, ix, 1 (March 2000): 41–46.

[9] Coliva, "Thought Insertion," p. 28.

[10] In this paper "the sense of agency" and "agency" are used interchangeably. Both terms refer to first-person, conscious experience. For recent discussion of "agentive experience" see Tim Bayne, "The Sense of Agency," in Fiona Macpherson, ed , *The Senses* (New York: Oxford, 2011). Also see below, note 63

I. SHOEMAKER'S IMMUNITY PRINCIPLE

In his reflections on self-consciousness, Shoemaker takes as a point of departure what he regards as an incontrovertible conceptual truth: "an experiencing is necessarily an experiencing by a subject of experience."[11] He evinces that a subject and an experience are just as intimately related as are a branch and a branch-bending. He then proceeds to develop a conception of self-consciousness that aspires to compatibility with both naturalism and certain Cartesian intuitions.

Developing one among these intuitions, and taking his lead from Wittgenstein, Shoemaker marks a distinction between the use of "I" (and its cognates) "as subject" and its use "as object."[12] Use-as-subject refers to such expressions as "I am in pain"; use-as-object refers to such expressions as "I am bleeding." Imagine, for example, that a base-runner and a catcher collide at home plate. As is not uncommon, the catcher's leg is gashed by the spikes on the base-runner's shoes, although the catcher does not immediately feel any pain. Because their limbs are entangled, upon first seeing the wound, the catcher does not immediately recognize it as his. As they disentangle, and as the catcher notices distinguishing features like the differences in their uniforms, he comes to realize that it is he who is bleeding. Recognition from the outside, so to speak, as in identifying the source of the bleeding, is recognition of self-as-object. The experience of pain, by contrast, given that it is known through introspection, typifies knowing about the self "as subject."

Wittgenstein's guiding intuition, one which is endorsed by Shoemaker, is: "...there is no question of recognizing a person when I say I have tooth-ache [*sic*]. To ask 'are you sure it is *you* who have pains?' would be nonsensical."[13] Shoemaker identifies the following as prototypical expressions of self-as-subject: "I feel pain"; "I am waving my arm"; and "I see a canary."[14] Clearly he believes that his argument is applicable to bodily sensation, to the sense of agency, and to perception of the external world. Take "I see a canary," for example: I might be mistaken concerning what I actually see (it might be a cardinal). I might even be hallucinating. But "it *cannot* happen that I am mistaken in saying this because I have misidentified as myself the person I know to see the canary."[15]

---

[11] Shoemaker, *First-Person Perspective*, p 10.

[12] Shoemaker, "Self-Reference and Self-Awareness"; Wittgenstein, *op. cit*, pp. 66–67.

[13] Shoemaker, "Self-Reference and Self-Awareness," p 556.

[14] *Ibid*, pp. 557–58. These three prototypical cases are all mental states that Shoemaker elsewhere describes as "weakly self-intimating". that is to say, it is the nature of these mental states to intimate themselves to their "possessors." Cf. Shoemaker, *First-Person Perspective*, pp. 50–52. Shoemaker adds that self-reference of this sort is not restricted to first-person pronouns: names and definite descriptions can self-refer in comparable ways. See Shoemaker, "Persons and Their Pasts," *American Philosophical Quarterly*, VII, 4 (October 1970) 269–85. The relevant discussion appears in footnotes 3 and 5. Also see Shoemaker, *Identity, Cause, and Mind*, p. 10, fn 4.

Why should it be nonsensical to query whether one is certain that it is oneself who is experiencing the mental state? Because, Shoemaker maintains, when we make a judgment like "I feel pain," we are aware that "one does, oneself, feel pain…one is, tautologically, aware, not simply that the attribute *feel(s) pain* is instantiated, but that it is instantiated *in oneself*."[16] Accordingly, it simply "*cannot* happen that I am mistaken in saying 'I feel pain' because, although I do know of someone that feels pain, I am mistaken in thinking that person to be myself."[17] The same is true for judgments about hand-waving or seeing canaries. Notice that these cases exude the modal force of "cannot." According to Shoemaker this is what makes self-consciousness special.

Shoemaker further elucidates IEM. He says that to claim that a statement "*a* is Φ" might be erroneous through misidentification relative to the term "*a*" is to allow for the following possibility: "the speaker knows some particular thing to be Φ, but makes the mistake of asserting '*a* is Φ' because, and only because, he mistakenly thinks that the thing he knows to be Φ is what '*a*' refers to."[18] But for IEM statements, mistakes of this type are not possible. If the ground of my judgment is introspection,[19] whenever I say "I feel pain" it cannot be the case that I am mistaken in thinking that the person in pain is me. Likewise, whenever I say "I am waving my arm" or "I see a canary" it cannot be the case that I have erroneously identified myself as the person who waves his arm or sees the canary.

How is it that immunity should obtain in such cases? Shoemaker replies that the relevant mental states are *identification-free*. He believes that even when we need to identify self (as-object), identification "will always presuppose the prior possession of other first-person information."[20] If self-consciousness always involved identification, whenever we self-ascribed a mental state (for example, "a is F") we would need to establish both "b is F" and "$a = b$." But "b is F," in turn, would further require that "c is F" and "$b = c$" be established. To avoid an infinite regress, we must allow for first-person knowledge that is not grounded in an act of identification.

---

[15] Italics added by the authors. Shoemaker distinguishes between absolute and circumstantial immunity. Our concern throughout this paper is exclusively with absolute immunity as regards the self-attribution of mental states.

[16] Shoemaker, "Self-Reference and Self-Awareness," pp 563–64.

[17] *Ibid*, p. 557. Also see Shoemaker, "Self-Knowledge and 'Inner Sense,' Lecture I. The Object Perception Model," *Philosophy and Phenomenological Research*, LIV, 2 (June 1994): 249–69, at p. 258; Shoemaker, *First-Person Perspective*, p. 15.

[18] Shoemaker, "Self-Reference and Self-Awareness," p. 557. Here we are only concerned with present-tense statements, but Shoemaker claims that IEM also holds for certain memory judgments. See Shoemaker, "Persons and Their Pasts."

[19] Cf. Pryor, "Immunity to Error through Misidentification," p. 279, and Joel Smith, "Which Immunity to Error?" *Philosophical Studies*, CXXX, 2 (August 2006): 273–83.

[20] Shoemaker, "Self-Knowledge and 'Inner Sense,' Lecture I," p. 258.

To illustrate this concern, consider the following. If I notice someone on a shopping center video display, I might wonder whether it is me. In order to make a proper identification, I might pull on my cap while checking to see whether the person on the video display does likewise. To perform this act of identification I must know that *I myself* am pulling on my cap. How can I know that? According to Shoemaker, my first-person knowledge that I am pulling on my cap must be grounded in identification-free first-person knowledge, because the only alternative would be just the sort of vicious infinite regress schematized above.[21]

Identification-freedom is also integrally related to his views on introspection, the self-intimating character of mental states, and the impossibility of self-blind creatures. For Shoemaker, introspective knowledge refers to just routine, mundane sorts of knowledge.[22] In his reflections on how best to understand introspection, he rejects "inner sense" models, notably the "object perception model" (OPM) and the "broad perceptual model" (BPM).

According to Shoemaker, if OPM is correct, then "identification information" about the perceived object must be available.[23] Critically, these objects would need to be independent of acts of perception. But Shoemaker denies that there is any such role for awareness of self-as-object to play in the explanation of introspective knowledge. Although it might appear to be the case that self is a good candidate for being an object of perception, Shoemaker believes that when we do need to identify self-as-object, identification "will always presuppose the prior possession of other first-person information."[24] Again, the only alternative to freedom from identification would be profligate identification, identification that cannot but lead to vicious infinite regress.

Shoemaker also rejects BPM, which differs from OPM in concerning itself with facts rather than objects.[25] Despite this difference, though,

---

[21] Shoemaker, "Self-Reference and Self-Awareness," p. 561, and *First-Person Perspective*, p. 196

[22] "The knowledge I have in mind is…the humdrum kind of knowledge that is expressed in such remarks as 'It itches,' 'I'm hungry,' 'I don't want to,' and 'I'm bored.'" See Shoemaker, "Self-Knowledge and 'Inner Sense,' Lecture I," p. 249

[23] *Ibid.*, pp. 252–53.

[24] *Ibid.*, pp. 254, 258.

[25] Shoemaker, "Self-Knowledge and 'Inner Sense,' Lecture II: The Broad Perceptual Model," *Philosophy and Phenomenological Research*, LIV, 2 (June 1994): 271–90.

BPM shares a fundamental commitment to the view that in perception we have access to things that are independent of being perceived. So identification-freedom would be incompatible with this model too.

Shoemaker's rejection of BPM is also linked to his rejection of the possibility of "introspective self-blindness." He believes that a significant—and unacceptable—consequence of BPM is that it allows for the logical possibility of this particular kind of blindness.[26] To be introspectively self-blind with respect to certain kinds of mental phenomena would require that, despite being able to conceive of those phenomena (just as the blind can conceive of phenomena unseen), a creature would be unable to introspectively access them. According to Shoemaker, BPM is only worth taking seriously if self-blindness is regarded as a conceptual possibility.[27] But he regards this notion to be as absurd as the claim that we could have pains but be systematically and blithely unaware of them.[28]

In short, in addition to IEM, Shoemaker endorses a "modest Cartesianism," a "weak version of the self-intimation thesis" (WST). On this view, the existence of certain mental entities is constitutively related to their being available to introspection. For those mental states that have phenomenal character, for example, it is of their essence that having them "issues in the subject's being introspectively aware of that character, or does so if the subject reflects."[29] There might well be internal states to which we do not have introspective access, states that play an important role in causing behavior. But Shoemaker says such states would not count as mental. The proper way to think of the relationship between introspection and mental states is that "the reality known and the faculty for knowing it are...made for each other—neither could be what it is without the other."[30]

Most philosophers regard IEM as a semantic or conceptual thesis. Recall that, according to Shoemaker, when one proclaims self to be in pain "one does, oneself, feel pain...one is, tautologically, aware, not simply that the attribute *feel(s) pain* is instantiated, but that it is instantiated *in oneself*." Unlike Shoemaker, we do not regard this as a tautology. On the contrary, it can be subjected to empirical investigation. Our main thesis is: awareness that mental states are instantiated does not entail awareness that said states are instantiated in self. Unlike most critics of Shoemaker, for the sake of argument, we grant that most of his views are correct. Even so, we argue that genuine counter-examples to IEM exist.

---

[26] *Ibid.*, p. 273.

[27] Shoemaker, *First-Person Perspective*, p. 31.

[28] Shoemaker, "Self-Knowledge and 'Inner Sense,' Lecture II," p. 275.

[29] Shoemaker, "Introspection and Phenomenal Character," *Philosophical Topics*, XXVIII, 2 (2001): 247–73; cf. Shoemaker, *First-Person Perspective*, p. 31.

[30] Shoemaker, "Self-Knowledge and 'Inner Sense,' Lecture II," p. 289. Also see p. 275.

II. IEM AND BODILY SENSATIONS: SOMATOPARAPHRENIA

Somatoparaphrenia is a syndrome that is characterized by the sense of profound estrangement from parts of one's body.[31] It is typically found in patients who have suffered extensive right-hemisphere lesions (usually vascular).[32] Lesions in the temporoparietal junction are a common neural correlate of somatoparaphrenia, but deep cortical regions (for example, the posterior insula) and subcortical regions (for example, the basal ganglia) are also sometimes implicated.[33]

Somatoparaphrenia is also closely associated with proprioceptive impairment and often (not always) co-morbid with hemispatial neglect. Patients feel that a contralesional limb, most frequently the hand, seems not to belong to them; indeed, it often seems to belong to someone in particular, not uncommonly an acquaintance.[34] The sense of disownership can be so vivid that even after recovery patients continue to describe the estrangement in factive, not metaphoric, language.[35]

Baier and Karnath assessed the frequency of somatoparaphrenia's occurrence.[36] They recently examined 79, consecutively admitted, acute stroke patients with right brain damage. They found that 11 experienced estrangement: five exhibited asomatognosia, and six were afflicted with somatoparaphrenia. Of the six, two attributed ownership of the limb to their wives, three to their examining physicians, and one to a patient sharing the same room.

Somatoparaphenia is occasionally accompanied not only by hemispatial neglect, but also by tactile extinction (the loss of conscious tactile perception) in the estranged body part. Moro et al. demonstrated (for two cases) that merely by changing the position of the hands—moving the left hand across the midline of the body, over to the right-hand side—tactile sensation could be recovered.[37] Even though tactile sensation could be so readily recovered, the sense of limb disownership was unchanged.

---

[31] Giuseppe Vallar and Roberta Ronchi, "Somatoparaphrenia: A Body Delusion. A Review of the Neuropsychological Literature," *Experimental Brain Research*, cxcii, 3 (2009): 533–51.

[32] For an exceptional case, see Rogerio Beato et al., "Transitory Somatoparaphrenia Associated with a Left Frontoparietal Meningioma," *Journal of Neurology*, cclvii, 7 (July 2010): 1208–10. For competing explanations of apparent lateralization, see Gabriella Bottini et al., "Productive Symptoms in Right Brain Damage," *Current Opinion in Neurology*, xxii, 6 (December 2009): 589–98, at p. 591

[33] Vallar and Ronchi, *op. cit.*, p. 548.

[34] Instances characterized just by the sense of limb disownership are referred to as asomatognosia; instances wherein limb ownership is attributed to someone else—some specific person—are referred to as somatoparaphrenia. The two are anatomically distinct. See Todd Feinberg et al., "The Neuroanatomy of Asomatognosia and Somatoparaphrenia," *Journal of Neurology, Neurosurgery and Psychiatry*, lxxxi, 3 (2010): 276–81.

[35] Peter Halligan et al., "Unilateral Somatoparaphrenia after Right Hemisphere Stroke: A Case Description," *Cortex*, xxxi, 1 (March 1995): 173–82

[36] Bernhard Baier and Hans-Otto Karnath, "Tight Link Between Our Sense of Limb Ownership and Self-Awareness of Actions," *Stroke: Journal of the American Heart Association*, xxxix, 2 (2008). 486–88.

As regards our challenge to IEM, the most relevant case has been described by Bottini et al.[38] A woman ("FB") reported that her left hand belonged to her niece and that she, FB, felt no tactile sensations there. In FB's case the lesion was subcortical, involving the basal ganglia, white matter underlying the insula, as well as other areas. But, importantly, the primary somatosensory area—which is critical to processing tactile sensation—was preserved. As Bottini et al. record: "F.B.'s spared *ability to perceive* tactile stimuli, provided that these were referred to someone else's body, was evidently based on the survival of some elementary somatosensory cortical functions."[39]

In a series of controlled tests, FB, while blindfolded, was advised that the examiner would touch *her* left hand; next, the examiner would in fact touch the dorsal surface of FB's hand. Whenever this was done, FB said that she could feel no tactile sensations. When advised that the examiner was about to touch her *niece's* hand, however, upon being touched FB reported feeling tactile sensation. Here we begin to see the relevance of FB's case to IEM.

It should be borne in mind that FB was in all other aspects cognitively sound.[40] Moreover, in order to ensure that these tests would be reliable, catch trials—wherein FB was led to expect touches that were not forthcoming—were used. These trials were evenly distributed across three verbal warnings—I am going to touch your right hand, your left hand, and your niece's hand—and were administered in four sessions, two on one day, two on the next. It is of paramount importance to note that in not even one of 36 catch trials, nine each per session, did FB respond incorrectly.[41] In other words, when advised that she (or her "niece") would be touched, if no contact was made, FB always reported "no," no contact had been made.

---

[37] Valentine Moro et al., "Changes in Spatial Position of Hands Modify Tactile Extinction but not Disownership of Contralesional Hand in Two Right Brain-Damaged Patients," *Neurocase*, x, 6 (2004): 437–43.

[38] Gabriella Bottini et al., "Feeling Touches in Someone Else's Hand," *NeuroReport*, xiii, 11 (2002): 249–52.

[39] Italics added by the authors. Bottini et al., *op. cit.*, p. 251.

[40] According to the clinical report, FB was "fully oriented in time and space and did not show any other sign of mental deterioration on the Mini Mental State Examination (score: 26/31)." See *ibid*

[41] *Ibid.*, Table 1.

As we begin examining IEM in light of this case, let us assume, for the sake of argument, that Shoemaker's central theses are largely correct. WST is true; both OPM and BPM, false. Moreover, self-as-subject is indeed distinct from self-as-object.

But even if we grant to Shoemaker his principal theses, we are left with an explanatory puzzle: why is it that when FB is expecting to be touched (on the left hand), she feels nothing, whereas when she expects that her niece will be touched there, she is able to report tactile sensation? Why, despite the experimental controls in place (for example, blindfold and catch trials), is she able to judge that "her niece" has been touched? Typically, to say (a) "I am going to touch your arm," implies (b) "I am going to touch you." It would be nonsensical to say (a) without implying (b). Likewise, when the doctor says, "I am going to touch your niece's hand," she implies, "I am going to touch your niece." The concern here is not *where* the sensation will be felt, but *who* will feel the sensation. Pace the prototypical situations that motivate the Wittgenstein-Shoemaker intuition—it is *not* absurd to inquire as to who is the subject of experience.

Let us divide the experiment into two parts: FB expecting that she will be touched is Part 1. FB expecting that her niece will be touched is Part 2. FB's case should be regarded as directly relevant to IEM because she has been primed by the doctor to introspect. We argue that in Part 2 FB is misrepresenting her tactile sensation as belonging to someone else. It is not the case that FB is misrepresenting the location of a sensation, as, for example, the base-runner does if he first represents his own leg as bleeding and then discovers that the bleeding leg is attached to the catcher with whom he collided. Instead, in Part 2, from FB's first-person perspective, when introspecting on that tactile sensation she misrepresents herself, such that she is not the owner of the sensation. In short, FB commits an error through misidentification regarding just *who* is the subject of the sensation.[42]

---

[42] One might still worry that FB's error concerns the location, not the subject of experience. For example, one possible characterization of FB's Part 2 is "I feel the sensation in my niece's hand." One could then argue that the subject of experience is not misrepresented. We should not assume, however, that this spare description can do full justice to the perplexing phenomenology. One problem is that it implies that FB, in feeling the sensation, regards "her niece's hand" from a third-person point of view. Thus, it cannot fully capture the complex pathology. Why? Not only is the clinician baffled by the experimental results, FB too is baffled by her own paradoxical experiences (*ibid.*, p. 251). Both FB's experience and her relation to "her niece's hand" are first-personal, introspective. An appropriate characterization, therefore, must also capture the perplexity from the first-person perspective. A more scrupulous reconstruction of this pathological experience would be: "I am introspectively aware of my niece's sensation." Under this reconstruction, the subject of experience, or the ownership of sensation, can be misrepresented.

To repeat, we can concur with many of Shoemaker's central theses: (1) For every mental state there must be a subject who experiences it. Moreover, for the sake of argument, we can agree with Shoemaker's WST. Thus: (2) Every mental state is in principle available to introspection. And we think Shoemaker would be obliged to concede that FB can only have the experience of a tactile sensation in Part 2 by means of introspection.

Although Shoemaker does not explicitly adopt a position concerning the ownership of sensation, a natural interpretation of his views would be as follows. (1) and (2) conjoined suggest: (3) Every mental state is experienced by the one who is currently introspecting that state.[43] The position is made explicit by Coliva,[44] who takes herself to be "vindicating" Shoemaker's claim that "in being aware that one feels pain one is, tautologically, aware, not simply that the attribute *feels pain* is instantiated, but that it is instantiated in *oneself*."[45]

We have formulated (1)–(3) in a way that fully accommodates Shoemaker's views. Our argument is that (1)–(3) do not provide sufficient ground to establish IEM. Proponents of IEM fail to take into account that (1)–(3) do not imply: (4) Every mental state is, from the first-person point of view, *represented as* experienced by the one who is introspecting the state. It is (4) that is needed for IEM to hold. FB's case is a counter-example to IEM because (4) is not true of those cases for which FB is introspectively aware of tactile sensation in Part 2. Although the attribute *feels sensation* is instantiated from the first-person point of view, it is not the case that the tactile sensation is instantiated in self. FB does not represent it in that way. The two instantiations are not tautologically linked. For IEM to be true, (4) must hold necessarily. But it does not hold with strict necessity; hence, IEM fails.

It is then empirically possible for a subject, while introspecting a mental state (and thereby knowing that someone is undergoing that state), to be in error with regard to who is experiencing that particular mental state. Admittedly, this is counterintuitive. The

---

[43] Shoemaker does, however, imply this position. See "Self-Reference and Self-Awareness," pp. 559–60, 565–67. We could articulate (3) such that it even more completely reflects Shoemaker's preferred mode of expressing his position by adding two clauses: (a) introspective awareness of phenomenal character occurs "if the subject reflects," and (b) it is just such introspective awareness that enables us to render judgments of the sort under consideration here, such as "I feel pain" (or, as applied to this case, "I feel a tactile sensation"). The relevant passages from Shoemaker are quoted above: the citation for (a) is indicated in fn. 29 and for (b) in fn. 16. But neither (a) nor (b) is critical to our argument.

[44] Coliva, "Thought Insertion," pp. 28–29.

[45] Shoemaker, "Self-Reference and Self-Awareness," pp. 563–64.

Wittgenstein-Shoemaker intuition—it is absurd to inquire of the person who introspects and reports a toothache whether it is indeed that person who has the ache—strikes all of us as correct. But empirical inquiry has ways of upsetting the apple cart: it would by no means be absurd to ask of FB whether it is she who has the tactile sensations, even though it is she who produces the introspectively based report.

An important contrast here calls for explanation. We have a fact and a foil,[46] the contrast between the two parts of FB's case. In Part 1, when FB is primed to introspect on what she experiences, she reports nothing; in Part 2, when she is primed to introspect on what her niece experiences, she reports tactile sensation.

To ignore this difference would be to ignore a significant explanatory problem. Because Parts 1 and 2 have similar histories, it is possible to ask sensible contrastive questions that enable us to elicit causal differences.[47] This possibility is not permitted by IEM. The essential difference between the two parts is whether FB represents herself as the subject of the mental state. This issue, concerning first-person representation of just *who* the subject is, we refer to as *mental ownership*.

One might worry that FB merely reports feeling the sensation, when in fact she does not feel anything. But had there been no actual sensations on this series of tests, the reports would not have been made. First, recall that in FB's case the lesion was subcortical; her somatosensory cortex was preserved. So it is not surprising that she retained the capacity for experiencing tactile sensations (provided that these were referred to someone else's body). Second, FB's performance on catch trials was perfect. Included among the catch trials were instances in which she was told that her niece's hand was going to be touched, when in fact it was not touched. In these trials she never once made a false report—neither on the first nor the second day of the experiments. Third, other standard procedures were in place to monitor FB's sustained attention and reliability of response: FB's hand was touched in a randomized, fixed sequence, which was repeated in four sessions on two separate days. Because she was blindfolded and because of the other controls that were in place, she could only have given an accurate report had she actually experienced the sensation of being touched.

Moreover, imaging studies of self-referential processing show that there is no reason to suspect that problems pertaining to mental ownership typically impair tactile processing. Northoff et al.'s analysis of many and varied studies that engage the "feeling of mineness"[48] indicates that these experiences are subsumed by a set of commonly activated regions within the cortical midline structure (CMS), regions that do not include the somatosensory cortex.[49] More specifically, as regards somatoparaphrenia, Feinberg et al., in a detailed study of 13 patients who had been examined using brain-imaging techniques within one week of acute hospitalization, identified a pattern of lesions distinctive of those who exhibited its clearest symptoms: repeated, refractory expressions of the conviction that their limbs belonged to someone else.[50] In this study, the region found to be most distinctive was not the somatosensory cortex; rather, it was the orbitomedial frontal cortex.[51] The claim here is not that any one region of the brain plays an exclusive, causal role in the etiology of somatoparaphrenia. The claim is that there is no empirical reason to suppose that what underlies the distinctive phenomenology of somatoparaphreia necessarily involves incapacitation of areas critical to somatosensory processing.[52]

---

[46] See Peter Lipton, "Contrastive Explanation," in David-Hillel Ruben, ed., *Explanation* (New York: Oxford, 1993), pp. 207–27.

[47] *Ibid.*, pp. 217–19.

The only remaining reason to suspect that FB actually did not experience the sensations is the worry that her case is analogous to blindsight. In the case of numbsense—also referred to as "blindtouch"—although subjects lack conscious tactile experience, they are to some degree capable of nonconsciously processing tactile information.[53]

---

[48] Georg Northoff et al., "Self-Referential Processing in Our Brain—A Meta-Analysis of Imaging Studies on the Self," *NeuroImage*, xxxi, 1 (May 2006): 440–57. In characterizing the "feeling of mineness" the authors say that "the self we consider here is an experiential self that mediates ownership of experience." See p. 441.

[49] Concerning the role of the CMS, Northoff et al. say, "Taken together, our results suggest that self-referential processing is mediated by cortical midline structures....We conclude that self-referential processing in CMS constitutes the core of our self and is critical for elaborating experiential feelings of self, uniting several distinct concepts evident in current neuroscience." See p. 440, also see pp. 448–49. As regards the relation between CMS and sensory processing, they report "Our review of neuroimaging studies reveals a set of commonly activated regions, within the extended CMS, during self-related tasks using a diverse set of sensory modalities. Activation in CMS must therefore be considered independent of the sensory mode within which the self-related stimuli were presented. Such sensory independence of neural activity in CMS can be observed in all domains." See p. 449.

[50] Feinberg et al., *op. cit.*

[51] "The orbitofrontal lesion was critical in the development of somatoparaphrenia versus simple asomatognosia." See *ibid.*, pp. 279–80.

[52] Cf. many of the cases reviewed by Vallar and Ronchi, *op. cit.*, Table 1, pp. 536–37.

[53] See Jacques Paillard et al., "Localization without Content: A Tactile Analogue of 'Blindsight'," *Archives of Neurology*, xl, 9 (September 1983): 548–51; Alberto Gallace and Charles Spence, "The Cognitive and Neural Correlates of 'Tactile Consciousness': A Multisensory Perspective," *Consciousness and Cognition*, xvii, 1 (2008): 370–407; and Yves Rossetti et al., "Implicit Body Representations in Action," in Helena De Preester and Veroniek Knockaert, eds., *Body Image and Body Schema* (Philadelphia: John Benjamins, 2005), pp. 111–25

In other words, perhaps FB was informationally sensitive to being touched but not experientially sensitive to being touched.

However, FB's case could not have been an instance of numbsense. For one thing, well-studied cases of numbsense involve damage to the primary somatosensory areas, very much unlike the case of FB. More importantly, in cases of numbsense the ability to make verbal report is lost. The reason given for ascribing numbsense is that the patients are able to point, with a moderate degree of accuracy, to the place where they were touched. In other words, by analogy to blindsight, their success at guessing based on nonconscious information is indicated by pointing, not by verbal report. FB's case is clearly not like this, since her capacity for verbal report is well preserved.

In conclusion, it seems that we are not immune in the way that IEM indicates. FB's introspections give rise to puzzling responses that are not compatible with IEM. Shoemaker's critical mistake might have been to infer from "what can happen as a matter of course," to what must necessarily be true of introspection and mental states.[54]

III. IEM AND THE SENSE OF AGENCY: BODY SWAP ILLUSION

The case against IEM can be made in multiple ways. In the previous section we dealt with bodily sensations. Here we show that similar problems can arise for the sense of agency concerning mental ownership.

Cognitive neuroscientists have been investigating whether specially designed experiments can induce in healthy subjects certain illusions pertinent to bodily self-consciousness. For example, in the case of the "Rubber Hand Illusion" it has been shown that ordinary people can experience an artificial hand as their own.[55] In these experiments, investigators primarily have been interested in the ownership of body parts and various phenomena that involve self-as-object rather than self-as-subject. In at least some of these experimental cases, however, issues relevant to IEM and self-as-subject are implicated. Most noteworthy among these is the "Body Swap Illusion."

In this case the illusory experience of owning a body that belongs to someone else is induced in healthy subjects. Although some among the neuroscientists who discovered the illusion are concerned only with body ownership, we argue that some of their experiments actually involve ownership of mental states. In the particular case described below, subjects can misrepresent themselves as experiencing someone else's experiences. After describing the experiment, we explain how it violates IEM.

---

[54] Shoemaker, "Self-Knowledge and 'Inner Sense,' Lecture II," p. 273.

[55] Matthew Botvinick and Jonathan Cohen, "Rubber Hands 'Feel' Touch that Eyes See," *Nature*, CCCXCI (February 19, 1998): 756.

The Body Swap Illusion was first demonstrated in a series of experiments conducted by Petkova and Ehrsson.[56] In one setting (their Experiment 5), two persons were involved: experimenter and subject. The experimenter wore a helmet outfitted with two closed-circuit television (CCTV) cameras, which transmit signals to a specific place. By positioning the cameras thus, the scenes they registered presented the experimenter's viewpoint. Wearing a set of head-mounted displays (HMDs), the subject stood face to face with the experimenter. The subject's HMDs were connected to the two CCTV cameras on the experimenter's head such that the images from the CCTV cameras played on the HMDs. The effect of this set-up was that the subject, adopting the experimenter's perspective, visually perceived himself rather than the experimenter.[57] The subject could see his own body from the shoulders to slightly above the knees. Both experimenter and subject were instructed to extend their right hands and then take hold, as if to shake. During the course of the experiment the two were instructed to squeeze one another's hands repeatedly, each time for two minutes. In the illusion condition, they squeezed in a synchronous manner; in the control condition, they squeezed asynchronously, alternating, the experimenter responding to the subject semi-randomly.[58]

Twenty subjects participated in this experiment, and each was interviewed immediately afterwards. The authors claim that the experiment "demonstrated that this set-up evoked a vivid illusion that the experimenter's arm was the participant's own arm and that the participants could sense their entire body just behind this arm."[59] To obtain more objective, quantifiable data, the scientists incorporated an anxiety-inducing threat into the experimental design (a knife above the wrist to suggest cutting of the hand) and measured each subject's skin conductance response (SCR). They reported that they "observed significantly stronger skin conductance responses when the knife was moved near the experimenter's wrist than when it was moved towards the participant's own hand in the synchronous condition."[60]

This experiment has significant implications for IEM. Note that in describing the participants' phenomenology, the authors say, "after the experiment, several of the participants spontaneously remarked...'I was shaking hands with myself!'"[61] Although the subjects "could clearly recognize themselves and distinguish between their own arm and the arm of the experimenter," this illusion is so robust that "a participant can face his or her biological body and shake hands with it without breaking the illusion."[62]

---

[56] Valeria Petkova and Henrik Ehrsson, "If I Were You. Perceptual Illusion of Body Swapping," *PloS One*, III, 12 (December 2008): 1–9.

[57] Recall that one of Shoemaker's three prototypical examples of introspection is "I see a canary." See "Self-Reference and Self-Awareness," p. 557.

[58] Petkova and Ehrsson, *op. cit.*, p. 4, Figure 6.

[59] *Ibid*, p 5.

[60] *Ibid*, Figure 7.

How should this aberrant experience be understood? The most natural way to characterize the subjects' phenomenology is with respect to agency. When they experience the illusion of shaking hands with themselves, their experiences involve misrepresentation of action awareness—that is, the misrepresentation concerns "who" shakes their hands. This poses a problem for Shoemaker's IEM.

From the subjects' first-person point of view, the handshaking experience involves the awareness that I am the agent of this action. This recently has been called "agentive experience" or "agentive self-awareness"—I experience myself as someone who is doing something.[63] What is distinctive about the Body Swap Illusion is that the subjects' agentive experience is mistaken. Although it was really the experimenter who was shaking their hands, the subjects misrepresented themselves as the agent of the action.[64]

To see how this creates a problem for Shoemaker's IEM as regards the case in question, we can agree with Shoemaker on each of the following: (1) For every agentive experience there must be a subject who experiences it. (2) Every agentive experience is in principle available to introspection. (3) Every agentive experience is experienced by the one who is currently introspecting it. The crucial point, however, is that (1)–(3) together do not imply: (4) Every agentive experience is, from the first-person point of view, *correctly represented as* experienced by the one who is introspecting it. Without (4) the ground upon which IEM stands is shaken.

Recall that one of Shoemaker's prototypical examples of self-as-subject is "I am waving my arm." According to IEM, it cannot happen that I am mistaken in saying "I am waving my arm" because although I do know of someone that is waving his arm, I am mistaken in thinking that person to be myself. I am necessarily aware that I am, myself, waving my arm. Mutatis mutandis for handshaking. This too clearly involves agentive experience. But as the body swap case shows, having an experience of handshaking does not guarantee that the agentive experience cannot be misrepresented. The mode of representation matters. Here, while in the illusory state, I am introspectively aware of the shaking hands, but I misattribute agency. I commit an error that violates IEM.

---

[61] *Ibid.*, p. 5.

[62] *Ibid.*, p. 1.

[63] Tim Bayne provides a clear example of agentive experience: "It's your first day as a waiter/waitress, and you are pouring water into a glass from a jug. As you pour the water, you experience yourself as an agent. You experience yourself as someone who is doing something, rather than someone to whom things are merely happening." See Bayne, *op. cit.*, p. 3. Also see Bayne and Elisabeth Pacherie, "Narrators and Comparators: The Architecture of Self-Awareness," *Synthese*, CVIX, 3 (December 2007): 475–91.

[64] Bodily sensations might play a contributing role: light pressure, intensity, duration, and location on the hand, at the points where the hand is squeezed, can all be experienced. But the illusion—"I was shaking hands with myself"—primarily involves action awareness. Also, as we have argued in section II, IEM can be violated in the case of bodily sensations.

In the case of somatoparaphrenia, a subject violates IEM because she experiences a mental state in virtue of having represented that state as belonging to her niece. In the case of body swap, subjects violate IEM because they represent themselves as agents when plainly they are not. In both cases IEM is violated. Introspective awareness that a mental state is instantiated, pace Shoemaker, does not prevent us from error as regards mental ownership.

IV. RESPONSE TO POSSIBLE CRITICISMS

In this section we consider three possible objections. The first concerns the relationship between conscious experience and reportability. The second concerns whether IEM should be regarded as a conceptual truth, and the third concerns an alleged distinction between agency and ownership.

First, the reason FB's case is particularly troubling for IEM is that it consists of two parts which reveal an explanatory contrast. In Part 1, when told that she will be touched, FB does not feel the sensation; yet, in Part 2, when told that her niece will be touched, she feels the sensation. The contrast exhibited here provides strong support for the claim that the self-as-subject of the relevant mental state is misrepresented in Part 2. Since FB felt the tactile sensation in Part 2, why didn't she feel it in Part 1? The only difference between Parts 1 and 2 concerns how the subject, from the first-person perspective, represents with regard to *who* is to be touched.

To salvage IEM, one might consider an alternative interpretation of her responses. Perhaps FB actually felt the sensation in Part I but, due to her pathologies, just could not report them. Were this so, the critical issue posed by this case might turn on the ability of FB to report tactile phenomenology, not the phenomenology itself. Proponents of IEM then could argue that IEM remains unchallenged because it does not presuppose a necessary connection between reportability and phenomenology. They could argue that FB felt the sensations both in Part 1 and Part 2, so it did not really matter *whom* the doctor said would be touched. It was just that FB failed to report it in Part 1.

Successful defense of IEM, however, requires that this strategy not remain mere speculation. There must be some reason to suggest that it accurately describes what transpired in FB's case. But no well-motivated reason suggests itself. On the contrary, there is good reason to think our interpretation of FB's case is accurate. Once again, recall that in order to ensure the reliability of FB's reports, the doctors conducted several catch trials that were evenly distributed across three different prompts: your right hand, your left hand, and your niece's hand. When untouched, FB never reported any sensation. When her right hand was touched, she always, unerringly, reported sensation. When her left hand was touched, she never reported sensation. But when her "niece's" hand was touched, she recovered tactile sensation. There is simply no evidence to suggest that reportability was a problem for her.

A second possible criticism is related to Campbell's interpretation of schizophrenia.[65] Campbell takes IEM to be a datum in need of explanation.[66] Indeed, he acknowledges, in accord with Wittgenstein and Shoemaker, that people take for granted the absurdity of asking, "Someone has a headache, but is it me?" Nevertheless, he does suggest that "there is some structure in our ordinary notion of the ownership of a thought which we might not otherwise have suspected."[67]

Coliva criticizes Campbell's interpretation of schizophrenia, and it is her defense of IEM that might abet those who would argue against our position.[68] Here the concern is just what constitutes mental ownership. We should first emphasize that we do not agree with the entirety of Campbell's argument—neither is IEM a datum, nor is the Wittgenstein-Shoemaker intuition veracious. What we do share with Campbell is the contention that ownership, as it pertains to conscious experience, is more complex than typically is acknowledged.

Responding to Campbell concerning the ownership of mental states, Coliva contends, "If a subject is introspectively aware of pain, this just means that she is feeling pain…it is a matter of *conceptual truth* that if a subject is introspectively aware of a certain mental state, then she herself is having it and, therefore, *that mental state is her own*."[69] She emphasizes that, as a matter of conceptual truth, introspective awareness of a mental state guarantees that one is the owner of said state. In developing her position she contends that other than introspective awareness, there simply is no independent criterion for what is to count as ownership of a conscious state. She regards this as a "vindication" of Shoemaker's treatment of IEM as a tautology.[70]

---

[65] Campbell, "Schizophrenia, the Space of Reasons, and Thinking as a Motor Process," *The Monist*, LXXXII, 4 (October 1999): 609–25.

[66] Campbell, "Immunity to Error through Misidentification and the Meaning of a Referring Term," pp. 91–94; Campbell, "The Ownership of Thoughts," *Philosophy, Psychiatry and Psychology*, IX, 1 (March 2002): 35–39.

[67] Campbell, "Schizophrenia," p 610.

[68] Coliva, "Thought Insertion."

[69] *Ibid.*, pp. 28, 29. Original essay not italicized.

But Coliva's view cannot save IEM. To see why, for the sake of argument, let us agree that there really are no independent criteria. The problem then would be that lack of independent criteria would by no means imply that mental ownership cannot be misrepresented. The key statements (1)–(3), as discussed in the previous sections, can be reformulated as follows: (1) Every mental state belongs to a subject. (2) Every mental state is in principle available to introspection. (3) Every mental state belongs to the one who is currently introspecting that state. Notice that Coliva's view is fully accommodated by (3). As we have argued above, however, (1)–(3) together do not imply: (4) Every mental state is represented (from the first-person point of view) as belonging to the one who is introspecting the state. Coliva's objection fails because she neglects (4).

Furthermore, it is not clear that there are no independent criteria for determining mental ownership. Recall what transpires in the case of FB: when advised that she would be touched, she felt nothing. When advised that her niece would be touched, she felt tactile sensations. If we regard IEM as a tautology, if we believe that introspective awareness guarantees mental ownership, then we arbitrarily dismiss the possibility of discovering independent criteria. Such dismissal would be tantamount to begging the question. What has been discovered in the cases of somatoparaphrenia and the Body Swap Illusion is that first-person representation of the ownership of mental states does not comport well with what might seem to be logically necessary or conceptually guaranteed.

FB recovers sensation because she has been cued not to represent the touch as an experience of her own, but as an experience that belongs to her niece. In other words, how the subject represents the experience provides an independent criterion for determining mental ownership. There is no question as to whether or not it is FB who is providing a report based on introspection. So there is no denying that information concerning the tactile sensation is available to FB. But, from the first-person perspective, this is not the end of the story. Ownership of mental states is a more complex phenomenon than the received view of IEM allows.

---

[70] Shoemaker, "Self-Reference and Self-Awareness," pp 563–64.

To the self-as-subject, from the first-person perspective, it matters just how the relationship between self and sensation is represented. Call to mind that one of Shoemaker's projects has been to elucidate "the distinctive way mental states present themselves to the subjects whose states they are."[71] What we have found is evidence that the distinctive way mental states present themselves to subjects varies and that, for one form of representation, ownership is contentious. It seems we have cases for which it would be by no means idle or absurd to inquire as to whether the experiences of which a person is introspectively aware belong to that person.

A third possible objection pertains to Gallagher's distinction between agency and ownership.[72] He employs this distinction to deny that schizophrenic thought insertion causes problems for IEM, because he believes there is no doubt as to just where, experientially, these thoughts are. The patient might well be sincere in expressing the feeling that he is not the author of these thoughts, but that is not to deny that these thoughts occur within his stream of consciousness. "His judgment that it is *he* who is being subjected to these thoughts is immune to error through misidentification, even if he is completely wrong about who is causing his thoughts."[73] In short, although the patient disclaims authorship, he does not deny experiencing the thought. Even in schizophrenia, there remains a nontrivial sense in which the inserted thoughts belong to the patient. Accordingly, in view of the fact that the relevant error in the Body Swap Illusion concerns agentive experience, one might try to argue that Gallagher's distinction applies here as well.

There are several reasons why this defense of Shoemaker's IEM fails. (i) For those subjects who feel that they are shaking hands with themselves during the illusion, it would still be reasonable to ask Wittgenstein-Shoemaker questions: is it you who is having the experience of squeezing your own hand? Is it you who is shaking hands with yourself? Arguably the most compelling intuition that motivates IEM is that questions of this type are absurd. But here they are not absurd. This very fact—that these questions can be well motivated—indicates that IEM does not enjoy the kind of modal force claimed by Shoemaker.

(ii) Admitting misrepresentation of agentive experience is already detrimental to Shoemaker's IEM. Again, one of Shoemaker's prototypical expressions of self-as-subject is "I am waving my arm." As is the case with "I am in pain," "I am waving my arm" enjoys an absolute immunity. Unlike "I am in pain," though, here we have a clear instance of agency. Shaking hands, just like waving, implies agency. In other words, Shoemaker's elucidation of IEM would still be assailable.

---

[71] Shoemaker, *First-Person Perspective*

[72] Gallagher, *op. cit.*

[73] *Ibid.*, p. 231.

Recall Shoemaker's formal articulation of his claim: for a statement "*a* is Φ" to be erroneous through misidentification relative to the term "*a*" is to allow for the following possibility: "the speaker knows some particular thing to be Φ, but makes the mistake of asserting '*a* is Φ' because, and only because, he mistakenly thinks that the thing he knows to be Φ is what '*a*' refers to." But for IEM statements, mistakes of this type are not possible: whenever I say, "I am shaking hands," it cannot be the case that I am mistaken in thinking that the person who is shaking my hand is me. It cannot be the case that I have erroneously identified myself as the person who is shaking my own hand. But that is precisely what happens in the Body Swap Illusion.

(iii) Gallagher argues that the distinction he discerns in schizophrenic thought insertion is sufficient to rescue IEM. But schizophrenia is not analogous to the Body Swap Illusion. First, in the case of body swap the error does not concern a lack of agentive experience; instead, it involves the erroneous attribution of agency to oneself. In the illusory state, one takes credit for more—not less—than one is capable of. Violation of Shoemaker's IEM in this instance is not due to denial of agency.

Second, in describing the schizophrenic's attribution of agency, Gallagher observes, "with respect to agency, he is in a position to make only statements in which he uses the first person pronoun as *object*—and in such cases the immunity principle is not at stake, and therefore cannot be violated."[74] In other words, Gallagher's view is that because schizophrenics lack agentive experience, they can only regard the "author" of inserted thoughts as object. Schizophrenia, then, is conspicuously different from the case of body swap. In the latter case, the subject misattributes agency to self. That is to say, the agent is regarded not as object, but as *subject*. Since here the agent is a self-as-subject, Gallagher's distinction cannot safeguard IEM.

(iv) Finally, and decisively, this strategy simply would not work for the case of somatoparaphrenia. No parallel argument drawing upon Gallagher's distinction can be made. The somatoparaphrenia case involves no action on the subject's part whatsoever.

In previous sections, we argued that adequate explanation of somatoparaphrenia and the Body Swap Illusion requires recognition that the ownership of mental states can be misrepresented. In this section, we have responded to what we regard as the strongest defenses of IEM and contend that they are not successful. We therefore conclude that the best explanation of the relevant cases reveals that IEM, rather than being a conceptual truth, is an empirical hypothesis, open to verification or refutation. Indeed, this hypothesis is confronted by substantive counterexamples.[75]

---

[74] *Ibid.* Italics added by the authors.

V. CONCLUSION

We have, for the sake of argument, adhered to the distinction between self-as-subject and self-as-object. According to Shoemaker, "absolute" immunity applies to self-attribution of mental states only as regards the former. What we have discovered is that, even when concerned exclusively with self-as-subject, we are not *necessarily* immune to error in the way that Shoemaker claims.

According to Shoemaker, introspective awareness that one feels pain tautologically implies both that (a) the attribute *feel(s) pain* is instantiated and that (b) it is instantiated in *oneself*. But the cases examined here reveal that (a) and (b) are just contingently connected. It is important to distinguish between those states which are instantiated in someone and those states—of which one is introspectively aware—that are represented as belonging to oneself. Mental states can be introspectively available to a subject without being represented as owned by the subject.

Accordingly, even when considering self-as-subject, we are not immune to error through misidentification relative to the first-person pronoun. Misidentification is possible because we can represent the ownership of mental states variously. Because the ownership of mental states is surprisingly complex there is no guarantee that subjects will not misidentify the subject of experience.

---

[75] Thus far we have not discussed perception of the external world—the third of Shoemaker's prototypical cases But, in personal communication, Roland Zahn has described a recent case which suggests that even here IEM can be violated (Zahn is a Clinical Research Fellow with the Neuroscience and Aphasia Research Unit at the University of Manchester.) The patient was suffering from right inferior temporal hypometabolism, problems pertaining to the supply of or ability to metabolize glucose. Multiple clinical interviews designed to elicit the patient's phenomenology consistently revealed that visual experience required a two-step process: upon first becoming aware of an object it was not immediately obvious that the object was being seen by self. In order to recognize a visual experience as belonging to self it was necessary to take a second step. This step involved relating *what* was being seen to *who* was seeing it. Caution is warranted in interpreting this case, for although the condition persisted for at least two months, all information derives from clinical or diagnostic assessments No experimental protocols were employed. But if Zahn's description is accurate, then Shoemaker's IEM fails for all three of the prototypical cases.

Others who have considered the type of cases treated herein have given most attention to the disownership of body parts. But from the first-person perspective, the question as to whether one owns a body part is distinct from questions concerning the ownership of mental states. For example, in the Body Swap Illusion, subjects are able to distinguish their own arm from the arm of the experimenter. Nevertheless, the illusion that one is shaking one's own hand persists. Ownership of body parts does not necessarily imply ownership of mental states. By allowing for this possibility we are able to account for what otherwise would be a wholly baffling phenomenon: I can recognize the hand extended in front of me as belonging to someone else while simultaneously feeling that I am shaking my own hand.

What does it matter if the ownership of mental states is complex in the ways that we indicate? Although we do not separately develop the issue here, one implication seems to be that first-person mental states are not identification-free in the way that Shoemaker claims. And since identification-freedom is a linchpin for many of Shoemaker's views concerning mental states and introspection, its loss would betoken significant consequences for other aspects of his views on self-consciousness.

Most philosophers agree that the *what* of conscious experience can be misrepresented, but that the *who* can be misrepresented continues to strike many as absurd. Shoemaker's IEM is an articulation of this robust intuition, the intuition that is well expressed by Wittgenstein's rhetorical question. But once these intuitions are clearly articulated, they are better regarded as hypotheses. Otherwise they can be found to arrest growth in understanding. Failure to allow for possible misrepresentation of the subject of experience leads to failure to ask important questions in empirical contexts.

Shoemaker's articulation of IEM as a conceptual truth was an attempt to say what is distinctive about self-consciousness. As we have argued, however, IEM is neither datum, nor tautology, nor conceptual truth. It is a hypothesis. By showing that mental ownership can be misrepresented, we have exposed IEM's vulnerability. Progress in understanding self-consciousness will require further inquiry into the phenomenon of mental ownership.

TIMOTHY LANE

National Chengchi University

CALEB LIANG[76]

National Taiwan University

---

[76] The order of authorship was determined arbitrarily; this manuscript is completely collaborative.

THERE IS NO ARGUMENT THAT THE MIND EXTENDS

On the basis of two argumentative examples plus their 'parity principle', Andy Clark and David Chalmers argue that mental states like beliefs can extend into the environment. I raise two problems for the argument. The first problem is that it is more difficult than Clark and Chalmers think to set up the Tetris example so that application of the parity principle might render it a case of extended mind. The second problem is that, even when appropriate versions of the argumentative examples can be constructed, the availability of a second, internalist parity principle precludes the possibility of inferring that the mind extends. Choosing which parity principle we ought to wield would involve deciding beforehand whether or not the mind can extend. Thus, Clark and Chalmers beg the question by employing their parity principle rather than the internalist one. I conclude that they fail to provide a proper argument to support the extended mind thesis.

I first explain Clark and Chalmers (henceforth: C&C)'s argument for the extended mind thesis (henceforth: EM), which is based upon their argumentative examples concerning Tetris and Otto, plus their 'parity principle'. Next, I look closely at the Tetris example and show that it requires some tidying up before it has the right structure to operate in an argument for EM. This tidying completed, I present a serious problem that afflicts C&C's argument, even with the best versions of the argumentative examples in hand; in short, C&C's use of the parity principle begs the question against internalism. I conclude that although C&C can be credited with raising EM as an interesting hypothesis, they have not provided a well-motivated (as in non-question begging) argument for its truth. Internalists will see no need to take seriously an argument that builds the truth of externalism into its premises. What C&C do bequeath us, however, as we eventually will see, is an interesting dilemma concerning the scope of mentality.

I. THE ARGUMENT FOR THE EXTENDED MIND

C&C frequently express EM by saying that a person's mental states supervene not just on goings-on within her skull, but on some goings-on in the environment as well. But supervenience is clearly not a tight enough relation with which to express EM, nor to ground an argument for EM's characteristic claim that the mind extends beyond the head. To see this, imagine that our mental states 'supervened upon'

0022-362X/11/0802/100–108

God's whim—that His state of mind was determinative of our own. Alternatively, consider a substance dualist who maintains that the state of his mind supervenes on the state of his body.[1] In neither of these cases is it desirable to locate the supervenience base literally in the mind, and this prospect is made no more desirable just because the supervenience relation holds. So supervenience is neither here nor there as far as EM is concerned. C&C must mean something much stronger, it follows, when they ascribe mentality to processes outside of the organism. What they really intend, of course—as they elsewhere explicitly state—is the *realization* relation: the mind is held, on their view, to be partially realized in the environment.

As I construe it, C&C's argument that the mind is partially realized in the environment depends upon the application of their 'parity principle' to key examples, each of which features a pair of functionally isomorphic states or processes, one existing inside the organism, the other outside.[2]

In their Tetris example, C&C compare the shape rotation that a conventional Tetris player achieves manually on a videogame screen with imaginative visualization of the same rotation. In their Otto example, C&C compare the storage of endorsed content in a notebook with the storage of endorsed content in the cranial biological memory. What makes it plausible that the internal and external resources in each case are functionally isomorphic is the fact that the agent concerned achieves his end just as well and more or less just as easily via the use of either one. Tetris can be played just as well and more or less just as easily on a Gameboy screen as in the head, and Otto can locate the Museum of Modern Art in New York just as well and more or less just as easily by consulting a notebook as by consulting his onboard memory.

Given that the internal resources in the argumentative examples—mental rotation of shapes and cranial storage of standing beliefs—are considered by C&C to be unproblematically mental phenomena, they utilize the apparent functional equivalence of these resources with their external analogues in order to stretch ascriptions of mentality out into the environment. They achieve this result through the application to the examples of what they call *the parity principle*

---

[1] David Chalmers considers such a position, for example, under the rubric 'naturalistic dualism'. See his *The Conscious Mind* (New York: Oxford, 1996).

[2] The argument is first proposed in Andy Clark and David Chalmers, "The Extended Mind," *Analysis*, LVIII, 1 (Janurary 1998): 7–19. See also Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (New York: Oxford, 2008), for the latest formulation and defense of EM, which clarifies the original argument structure. It is Clark's reworking that provides my main target here.

(henceforth: PP), expressly intended, says Clark, to free us from our "biochauvinistic prejudice"[3]—our unjustified tendency to ascribe mentality exclusively to *intra*-organism processes:

> PP: If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive/mental process, then that part of the world is (for that time) part of the cognitive/mental process.[4]

On this basis, assuming rough functional equivalence between internal and external resources plus the parity principle, C&C conclude that manually rotating tiles on a videogame screen and storing accessible and endorsed content in a notebook are cognitive/mental processes, that is, cases of the mind's partial realization in the environment.

II. TIDYING UP THE TETRIS EXAMPLE

There is an important difference between mental rotation of Tetris shapes and the manual rotation one achieves on a videogame screen, a difference that threatens to wreck C&C's use of the Tetris example. Mental rotation of Tetris shapes is, notably, a *conscious* process: indeed, its intrinsically conscious self-presentation to the subject will be essential for conducting her imagination-based game of Tetris. Roughly speaking, this playing subject visually imagines the Tetris tiles rotating in order to identify them, assess their prospects for integration into the growing wall, and then (mentally) place them appropriately. When playing Tetris on a Gameboy, on the other hand, the screen is not a resource of which the player is immediately aware; she is not directly conscious of this representation of the game. The representation of the Tetris game on the Gameboy screen is one thing, this player's conscious experience of the game, another. If we take PP at its strict letter, therefore, the Tetris example gets nowhere near to establishing EM: for it is not the case that, were something like a Gameboy running in the Tetris player's head, this would by itself help constitute a cognitive/mental process. Simply having a videogame going on inside one's skull hardly counts as cognitive/mental activity![5] The external process in this case is not such that were it occurring in the head we would count it as cognitive/mental. PP does not, therefore, recommend counting a Gameboy on which one is playing as a cognitive/mental resource.

---

[3] Clark, *op. cit.*, p. 77. In fact, in the original paper PP is introduced on the basis of *first* considering the Tetris example. Nonetheless, Clark and Chalmers make it clear that it is PP that (also) drives their reasoning about the Tetris case.

[4] Slightly adapted from Clark and Chalmers, *op. cit.*, p. 8. In the original PP the term 'mental' does not figure: C&C took the Otto example to establish extended mentality (because of the assumed mentality of internal standing beliefs), and the Tetris example to establish EM to the extent that cognition implies mentality, for they take this example to establish extended cognition at least. I set aside this difference between the examples here, as well as the question of any shortfall between the cognitive and the properly mental; nothing substantive hangs on these issues for my present purposes.

In their two cases intermediary between uncontroversially *mental* (imaginative) shape rotation and conventional (external) onscreen Tetris play, C&C first envisage a Martian with a biologically inherited motor-controlled retinal display, and then a cybernetically engineered near-future human with a similar retinal display who can choose between using this or employing mental visualization to rotate Tetris shapes. C&C claim that the Martian's inherited apparatus is indisputably a cognitive/mental resource and that this encourages the same classification of the technological endowment of the near-future human. By functional parity—via PP—therefore, the shape rotation a conventional Tetris player accomplishes on an external Gameboy screen also counts as cognitive/mental activity, activity realized in the extra-cranial processes of the Gameboy. In light of the difficulty we identified above for C&C's claim that conventional Tetris play can constitute an instance of EM, it is worth lingering over the two intermediate cases to see whether we might glean from them a clearer idea of how to read the Tetris argumentative example in its strongest possible form.

Considering the intermediate cases further, we can see that there is some awkwardness in the fact that the internal display-screens are located at the retina. For it is not obvious that the retina counts as a cognitive/mental resource; when one has a scratched retina, for instance, it does not seem to follow that one has a scratched cognitive system/mind. Of course, *if* the mind can extend then perhaps the retina could become part of it, under the right circumstances. But that possibility cannot be assumed here without begging the question, for we are still in the stages of the argument for EM. This means trouble when we apply PP: were some environmental process to operate in a manner functionally isomorphic with the hypothesized retinal displays, this would not encourage us to think of that process as cognitive/mental, on account of the fact that its internal analogue is not cognitive/mental to start with. The external process would not be such that were something like it going on in the head we would count it as cognitive/mental. Hypothetical retinal displays are therefore a bad place to start if one wishes to build a C&C-style argument for EM.

---

[5] Of the right kind, anyway; it would surely have all sorts of unfortunate cognitive/mental effects.

This difficulty is easily overcome, however. We need only recast C&C's retinal display as a display-screen realized somewhere further inside the brain, in an uncontroversially cognitive/mental system—within the visual cortex, for example. Having taken the screen away from the eye, though, we now face the crucial question about the subject's awareness of what it represents: are we to imagine the screen set up so that (somehow) the subject is directly conscious of what is on it, so that the screen in some way *constitutes* the player's conscious awareness of the Tetris game? Or are we to imagine things so that the screen merely *causes* (not constitutes) the player's conscious visual representation of the game? (For the latter interpretation, picture a Gameboy screen embedded in the player's skull and then suitably patched into her visual system,[6] so that she experiences just the same visual imagery of the Tetris game as she does when holding a Gameboy close to her face.)

The constitutional reading creates a new problem for C&C's application of PP in support of EM, for it opens up a new discrepancy between the internal and external resources. It is not the case that, were a Gameboy implanted (however ingeniously) into someone's head, it would constitute their conscious representation of a Tetris game. On the constitutional reading, therefore, even if the Martian/near-future human's internal display is a cognitive/mental resource, PP does not recommend also counting a normal (external) Gameboy as cognitive/mental: for the latter is not such that were it running in the head we would consider that a cognitive/mental process. Such an internal process would need to be a conscious representation of the Tetris game, but a Gameboy screen is not itself a conscious representation of anything, wherever it is.

It follows that C&C must—or at least ought to—intend the causal reading of the Martian/near-future human cases. That is, we are to think of the internal displays as causing these Tetris players' conscious representations of how things are with their games. Now we seem to have achieved the desired internal/external functional isomorphism, for a normal Gameboy in play is also the cause of the player's conscious representation of the game she is playing. We have succeeded in tidying up the Tetris argumentative example: as long as the internal cause of a playing subject's conscious representation of her Tetris game counts as a cognitive/mental resource, PP will encourage us to count an external cause of such a conscious representation—something like a Gameboy—as a cognitive/mental resource as well. In this way EM will be supported.

---

[6] Plus motor cortex, as before, so that she can control the game.

However, this optimal version of the Tetris example faces a serious objection that applies equally to the Otto example, to which we turn next.

III. OTTO, TETRIS, AND PARITY PRINCIPLES

The difficulty for C&C's use of the Otto example is that so-called 'parity principles' are, it turns out, relatively cheap and easy to come by. Here is a second one:

> PP*: If, as we confront some task, a part of the head functions as a process which, were it to go on in the world, we would have no hesitation in rejecting as part of the cognitive/mental process, then that part of the head is (for that time) not part of the cognitive/mental process.

This principle yields the conclusion that standing beliefs, stored in the organism, are not part of the subject's cognitive/mental process. For we do not *prima facie* regard content-storing states of the environment, whether accessible to consciousness or not, as mental. On the other hand, consciousness is clearly not disqualified as mental by PP*. Though we may consider it naturally impossible, there is little conceptual trouble in making sense of a subject's consciousness partially, or even fully, extending outside of her organism. This is why it is not *a priori* obvious that "out of body" experiences are illusory. In such a case we would accept the verdict that the subject's mental process had extended: thus consciousness passes muster according to PP*. PP* also makes the correct adjudication about processes such as the growth of myelin sheaths (that is, that they are not mental). Application of this principle to Otto does not establish EM; PP* would instead, applied broadly, help establish the thesis that Brie Gertler calls 'the narrow mind', the view that it is only occurrent consciousness that is properly speaking mental.[7] For it is, arguably, only where we find consciousness that we are compelled to posit mentality.[8] Otto would, on this reading, only have the content that MoMA is on 53rd Street 'in mind' when he reads his notebook and consciously entertains it.

Returning to the tidied Tetris example: the application of PP* here would also support an internalist conclusion, rather than EM. For it is not *prima facie* the case that an external screen that causes a conscious representation of a Tetris game in the player's mind—something like a Gameboy—counts as a cognitive/mental resource.

---

[7] Brie Gertler, "Overextending the Mind?" in Gertler and Lawrence Shapiro, eds., *Arguing about the Mind* (New York: Routledge, 2007), pp. 192–206.

[8] Gertler is far from alone in this controversial view of mentality; Galen Strawson, for example, is another important advocate. See Strawson, *Selves: An Essay in Revisionary Metaphysics* (New York: Oxford, 2009). I take no stand on the matter here; my key point in this paper is simply that nothing Clark and Chalmers say prejudices the issue against this kind of internalism.

By PP*, therefore, an internal analogue of such a screen would not count as cognitive/mental either. For it would be a process such that, were it going on in the environment, we would not count it as cognitive/mental. Let me be clear: this is not to deny, far from it, that internal *conscious* visualization of a Tetris game is a cognitive/mental process; indeed, I think most would agree that such a process is paradigmatically cognitive/mental. What PP* denies cognitive/mental status is, rather, any internal process merely *causally underlying* this kind of conscious visualization. For, to repeat, external causes of conscious representations are not themselves taken to be cognitive/mental phenomena.[9] On this reading of the Tetris example, neither the near-future human nor the Martian possesses genuinely cognitive/mental resources in the form of their internal Tetris displays.[10]

What the availability of PP* reveals is a disingenuous element in C&C's treatment: the parity principle they employ is not as even-handed as its name suggests; it is far from a "'veil of ignorance' style test" provided only to help us neutralize our organismic bias, as Clark suggests.[11] PP might appear harmless, because at a glance it seems to amount to no more than pointing out an equivalence between two types of item: an extra-cranial resource and an intra-cranial one. And equivalences do not have priorities or biases built in of themselves, we normally think. But it is also clear that, depending upon the interests of the person drawing a particular equivalence, it *is* possible to build priority into that equation; this hangs on which side of the equation the person takes as basic. The equation thereby can be imbued with directionality—bias. An analogous point is at the root of anti-physicalist complaints that physicalist reductionists in effect propose a kind of eliminativism about the mind, for example. Reductionists claim to be proposing no more than an equivalence when they say that mind 'just is' matter. But anti-physicalists find that when the material side of the equation is taken as basic, the explanandum that matters to them can seem to disappear, as if the victim of a conjuror's trick.

C&C also engage in conjuror's trickery with their use of PP. PP is ingrained with externalist bias; its application to the examples is designed to yield EM as the conclusion. I do not claim that PP* is any more innocent. (Note, however, that it can be given its own egalitarian-sounding gloss, as intended to help us avoid another variety of biochauvinistic prejudice: the tendency to think that if some resource is skull-based then it is *ipso facto* cognitive/mental. PP*, as much as PP, encourages us to ignore the boundary of skull and skin when it comes to attributing cognition/mentality, as Clark earnestly recommends.) Given its relation to PP, PP* wears its internalism on its sleeve; it is designed to reverse the verdict of EM and radically restrict the scope of mind.

---

[9] Except where we have independent reasons to consider them cognitive/mental, for example, when they are other people.

[10] When these are taken according to the causal reading of the scenarios. Recall that while on the constitutional reading the internal displays might count as cognitive/mental, they make functional isomorphism with an external Gameboy—and so, the application of PP—impossible.

[11] Clark, *op. cit*, p. 77.

PP and PP*, then, cancel one another out dialectically. We might at this point usefully introduce a third, second-order parity principle:

> PP^2: If, as we confront conflicting first-order parity principles, each principle equally can be given an egalitarian gloss, and each delivers equally on substantive but unobvious conclusions, then we should have nothing to choose between them *prima facie*.

In PP^2, I submit, we finally have a parity principle that operates without prejudice. It indicates that, given a choice of inconsistent first-order parity principles, we need some independent reason to endorse one rather than the other, in the sense of selecting it for application to argumentative examples of the sort we have been considering, and which C&C take to support EM. The question we now face, consequently, is this: why should we choose to employ PP instead of PP*, thereby arriving at C&C's externalism, rather than proceeding vice versa, by a path that will take us towards an equally radical internalism?

Does PP^2 provide us the means to decide between PP and PP*? I do not think so; by the criteria of PP^2 the two first-order parity principles are on a par. The conclusions they each support, that the mind can be partially realized in notebooks, Gameboys, and other environmental resources, and that the mind is limited to occurrent, conscious content (so that standing beliefs are nonmental) are equally surprising and radical, and from an impartial perspective each thesis departs equally far from the commonsense view of the mind (which I daresay has it both that the mind cannot extend and that standing beliefs are mental). If we are to resolve the issue of which parity principle to wield, then we will have to look for our reasons elsewhere.

IV. WHY THE ARGUMENT FOR THE EXTENDED MIND BEGS THE QUESTION

If we apply PP to C&C's argumentative examples, we arrive at EM. If on the other hand we apply PP*, standing beliefs turn out not to qualify as mental, and although imaginatively visualizing a Tetris game is undoubtedly a cognitive/mental activity, no relevantly analogous—that is, conscious—activity can occur outside of a head, as C&C

themselves concede.[12] PP* urges a much stricter conception of what qualifies as genuine mentality.

We need some means of deciding which first-order parity principle we ought to apply to the argumentative examples in order to know which way the argument for EM will go. However, I submit, there is simply no way to do this without sorting out in advance whether we think EM might be true or not. Those who will accede to applying PP to C&C's argumentative examples will be those antecedently sympathetic to the idea that mental processes might extend. Conversely, those with strong internalist leanings naturally will want to apply PP* to the examples. Further, if we think for a moment about the kinds of considerations which we might adduce to encourage someone to accept that it is, say, PP rather than PP* that should be applied to a given example, we see that these will be precisely considerations which already stand in favor of EM. And conversely. One will only find such considerations persuasive, it follows, to the extent that one is already disposed towards or against EM.

It seems that to decide which parity principle properly applies to a given argumentative example (consisting of a pair of functionally isomorphic putatively cognitive/mental resources, one internal and one external), one would have to have already made up one's mind about EM's truth. It follows that PP cannot be used in an argument for EM; rather, its adoption could only be the outcome of such an argument. Since C&C rely on the employment of PP, rather than a principle like PP*, to drive through their externalist conclusion, we must conclude that their argument for EM begs the question.

V. CONCLUSION

C&C have raised an interesting possibility, that the mind extends, but they have offered no serious (as in, non-question begging) argument in favor of the thesis. The real argument over EM, it seems, is yet to come for both sides in the debate. We end by noting a further interesting consequence of C&C's argument, given the options we now have on the table: it appears that we are forced towards either a radically externalist or a radically internalist conception of genuine mentality. Even if the correct choice (insofar as our concept of mentality admits of one) remains as yet undetermined, having uncovered this surprising dilemma would represent significant progress in the philosophy of mind, and for that C&C deserve credit.

SAM COLEMAN

University of Hertfordshire

---

[12] See, for example, Chalmers's foreword to Clark, *op. cit*

NEW BOOKS

AGAMBEN, GIORGIO: *Nudities.* Stanford: University Press, 2010. ix, 120 p. Paper: $16.95.

ALAIMO, STACY: *Bodily Natures.* Bloomington: Indiana UP, 2010. xi, 193 p. Paper: $24.95.

AXINN, SIDNEY: *Sacrifice and Value: A Kantian Interpretation.* Lanham, MD: Rowman & Littlefield, 2010. x, 143 p. Cloth· $60.00.

BAXLEY, ANNE MARGARET *Kant's Theory of Virtue: The Value of Autocracy.* New York: Cambridge, 2010. xvi, 189 p. Cloth: $85.00.

BAYNE, TIM: *The Unity of Consciousness.* New York: Oxford, 2010. xi, 341 p. Cloth: $55.00.

BECHTEL, WILLIAM, and ROBERT RICHARDSON: *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research.* Cambridge: MIT, 2010. liv, 286 p. Paper: $27.00.

BEINER, RONALD: *Civil Religion: A Dialogue in the History of Political Philosophy.* New York: Cambridge, 2010. xv, 432 p. Paper: $34.99.

BENHABIB, SEYLA: *Politics in Dark Times: Encounters with Hannah Arendt.* New York: Cambridge, 2010. ix, 397 p. Paper: $28.99.

BLOM, PHILLIP: *A Wicked Company: The Forgotten Radicalism of the European Enlightenment.* New York: Basic, 2010. xx, 361 p. Cloth: $29.95.

BRANDHORST, KURT: *Descartes' Meditations on First Philosophy.* Bloomington: Indiana UP, 2010. 213 p. Paper: $19.95.

BRAKEL, LINDA A. W.: *Unconscious Knowing and Other Essays in Psycho-Philosophical Analysis.* New York: Oxford, 2010. xii, 170 p. Paper: $82.95.

CAO, TIAN YU: *From Current Algebra to Quantum Chromodynamics.* New York: Cambridge, 2010. x, 308 p. Cloth: $85.00.

CASTOGNOLI, LUCA: *Ancient Self-Refutation: The Logic and History of the Self-Refutation Argument from Democritus to Augustine.* New York: Cambridge, 2010. xx, 394 p. Cloth: $99.00.

COHEN, JOSHUA: *The Arc of the Moral Universe and Other Essays.* Cambridge: Harvard, 2010. viii, 416 p. Cloth: $39.95.

DOUEIHI, MILDAD: *Augustine and Spinoza.* Cambridge: Harvard, 2010. ix, 116 p. Cloth: $29.95.

DWYER, JAMES G: *Moral Status and Human Life.* New York: Cambridge, 2010. viii, 212 p. Cloth: $90.00

ERNELING, CHRISTINA E.: *Towards Discursive Education: Philosophy, Technology, and Modern Education.* New York: Cambridge, 2010. viii, 202 p. Paper: $31.99.

FORSTER, GREG: *John Locke's Politics of Moral Consensus.* New York: Cambridge, 2010. x, 317 p. Paper: $32.99.

FRENCH, PETER A.: *War and Moral Dissonance.* New York: Cambridge, 2010. xv, 343 p. Paper: $30.99.

HYLTON, PETER: *Quine.* New York: Routledge, 2010. x, 405 p. Paper: $34.95

JOHNSON, DIRK R.: *Nietzsche's Anti-Darwinism.* New York: Cambridge, 2010. ix, 240 p. Cloth: $85.00.

LENUNISEN, MARISKA· *Explanation and Teleology in Aristotle's Science of Nature.* New York: Cambridge, 2010. xiii, 250 p. Cloth: $85.00.

LEVIN, ABIGAIL: *The Cost of Free Speech: Pornography, Hate Speech, and Their Challenge to Liberalism.* New York: Palgrave Macmillan, 2010. ix, 213 p. Cloth: $85.00.

MACKENZIE, ADRIAN: *Wirelessness: Radical Empiricism in Network Cultures.* Cambridge: MIT, 2010. 255 p. Cloth: $32.00.

MARRA, MICHAEL F.: *Japan's Frames of Meaning: A Hermeneutics Reader*. Honolulu: Hawaii UP, 2010. xv, 443 p. Cloth: $55.00.

MEISTER, ROBERT: *After Evil: A Politics of Human Rights*. New York: Columbia, 2010. x, 544 p. Cloth: $60.00.

MELE, ALFRED R.: *Effective Intentions: The Power of Conscious Will*. New York: Oxford, 2010. ix, 178 p. Paper: $24.95.

MURPHY, COLLEEN: *A Moral Theory of Political Reconciliation*. New York: Cambridge, 2010. viii, 214 p. Cloth: $85.00.

NAAMAN-ZAUDERER, NOA: *Descartes' Deontological Turn: Reason, Will, and Virtue in the Later Writings*. New York: Cambridge, 2010. xii, 224 p. Cloth: $85.00.

PELOSI, FRANCESCO: *Plato on Music, Soul and Body*. New York: Oxford, 2010. vi, 228 p. Cloth: $95.00.

PORTER, JAMES I.: *The Origins of Aesthetic Thought in Ancient Greece: Matter, Sensation, and Experience*. New York: Cambridge, 2010. xvii, 607 p. Cloth: $145.00.

RECANATI, FRANCOIS: *Truth-Conditional Pragmatics*. New York: Oxford, 2010. vii, 324 p. Paper. $35.00.

SANTAS, GERSAMIOS: *Understanding Plato's Republic*. Malden, MA: Wiley-Blackwell, 2010. xii, 238 p. Paper: $29.95.

SCHMIDTZ, DAVID, and JASON BRENNAN: *A Brief History of Liberty*. Malden, MA: Wiley-Blackwell, 2010. ix, 267 p. Paper: $29.95.

SCHUMACHER, BERNARD N.: *Death and Mortality in Contemporary Philosophy*. New York: Cambridge, 2010. xi, 258 p. Paper. $27.99

SIEGEL, SUSANNA: *The Contents of Visual Experience*. New York: Oxford, 2010. x, 222 p. Cloth: $40.00.

STEVENS, RICHARD G.: *Political Philosophy: An Introduction*. New York. Cambridge, 2010. xxi, 309 p. Paper: $27.99.

STEWART-WILLIAMS, STEVE: *Darwin, God and the Meaning of Life: How Evolutionary Theory Undermines Everything You Thought You Knew*. New York: Cambridge, 2010. x, 341 p. Cloth: $28.00.

STOCK, BRIAN: *Augustine's Inner Dialogue: The Philosophical Soliloquy in Late Antiquity*. New York: Cambridge, 2010. xiii, 240 p. Cloth: $85.00.

STUDMANN, PAUL: *Empiricism and the Problem of Metaphysics*. New York: Rowman and Littlefield, 2010. ix, 169 p. Cloth: $60.00.

TALBOTT, WILLIAM: *Human Rights and Human Well-Being*. New York: Oxford, 2010. xi, 410 p. Cloth: $85.00.

THOMMASON, AMIE L.: *Ordinary Objects*. New York: Oxford, 2010. xi, 240 p. Paper: $24.95.

WALL, JOHN: *Ethics in Light of Childhood*. Washington DC: Georgetown, 2010. ix, 204 p. Paper: $34.95.

WEITHMAN, PAUL: *Why Political Liberalism: On John Rawl's Political Turn*. New York: Oxford, 2010. xii, 379 p. Cloth: $65.00.

WILLIAMS, DAVID-ANTOINE: *Defending Poetry: Art and Ethics in Joseph Brodsky, Seamus Heaney, and Geoffrey Hill*. New York: Oxford, 2010. xii, 241 p. Cloth: $126.00.

WRATHALL, MARK S.: *Heidegger and Unconcealment: Truth, Language and History*. New York: Cambridge, 2010. xi, 250 p. Paper. $29.99.

WYNLTON, THOMAS: *On the Intellectual Soul*. New York: Oxford, 2010. lxiii, 140 p. $75.00.

YAFFE, GIDEON: *Attempts*. New York: Oxford, 2010. x, 349 p. Cloth: $75.00.

NEW FROM OXFORD

SOCIAL EPISTEMOLOGY
Essential Readings
Edited by ALVIN GOLDMAN and DENNIS WHITCOMB
The classic articles in this volume address the fundamental questions of social epistemology in ways that are both cutting-edge and easy to understand.
2011 368 pp. Hardback $99.00 Paperback $29.95

SELF-KNOWLEDGE
Edited by ANTHONY HATZIMOYSIS
This collection of specially commissioned essays enhances our understanding of the nature and the limits of knowledge of our own beliefs and desires.
2011 320 pp. $85.00

UNDERIVATIVE DUTY
British Moral Philosophers from Sidgwick to Ewing
Edited by THOMAS HURKA
A team of eminent contemporary philosophers present the first collective study of seminal British moral thinkers of the late 19th and early 20th centuries.
2011 280 pp. $65.00

SAVORING DISGUST
The Foul and the Fair in Aesthetics
CAROLYN KORSMEYER
This book is a fresh approach to the old problem of why emotions that are considered "negative" when experienced in practical life are often valued and sought after in art.
2011 208 pp. Hardback $99.00 Paperback $29.95

OCCASIONALISM
Causation Among the Cartesians
STEVEN NADLER
Steven Nadler presents a collection of essays on philosophical problems about causation in the seventeenth century.
2011 256 pp. $65.00

PROMISES AND AGREEMENTS
Philosophical Essays
HANOCH SHEINMAN
This volume, which comprises sixteen original contributions, is the first collection of philosophical papers on promises and agreements.
2011 432 pp. $65.00

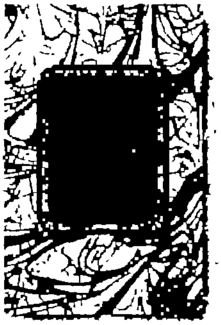
DAVID HUME: A TREATISE OF HUMAN NATURE
Two-volume set
Edited by DAVID FATE NORTON and MARY J. NORTON
David and Mary Norton present the definitive scholarly edition of Hume's *Treatise*, one of the greatest philosophical works ever written. This set comprises the two volumes of texts and editorial material, which are also available for purchase separately.
2011 1200 pp. Quantity Pack $75.00

THE CONTENTS OF VISUAL EXPERIENCE
SUSANNA SIEGEL
In this book, Susanna Siegel develops a framework for understanding the contents of visual experience, and argues that these contents involve all sorts of complex properties.
(Philosophy of Mind Series)
2011 240 pp. $49.95

COLLECTED PAPERS, VOLUME 1
Mind and Language, 1972-2010
STEPHEN STICH
This is the first of two volumes collecting articles by the distinguished philosopher Stephen Stich. This volume collects the best and most influential essays that Stephen Stich has published in the last 40 years on topics in the philosophy of mind and the philosophy of language.
2011 416 pp. $65.00

Prices are subject to change and apply only in the US.
To order or for more information, visit **www.oup.com/us**.

OXFORD
UNIVERSITY PRESS



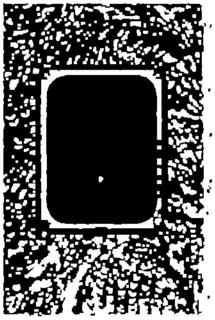
Hegel on Self-Consciousness

Desire and Death in the *Phenomenology of Spirit*

Robert B. Pippin

"*Hegel on Self-Consciousness* is the best treatment of the subject in English, French, or German—it is also clear, short, and to the point. It will very likely serve as a basic text in a variety of undergraduate and graduate courses. Pippin is one of the most widely read philosophers of our time."
—Terry P. Pinkard, Georgetown University

Princeton Monographs in Philosophy
Harry G. Frankfurt, Series Editor
Cloth $29.95 978-0-691-14851-9



Michael Oakeshott's Skepticism

Aryeh Botwinick

"The literature on Oakeshott tends to emphasize the independence of his thought and the problem of how to classify him, a problem that Aryeh Botwinick neatly sidesteps by taking an entirely different line. Oakeshott turns out to be in agreement or dialogue with a range of other thinkers, including Nietzsche and Arendt. This is a fresh and rewarding approach that provides a new perspective on Oakeshott's thought and its significance."
—Oliver Leaman, University of Kentucky

Princeton Monographs in Philosophy
Harry G. Frankfurt, Series Editor
Cloth $35.00 978-0-691-14717-8

Knowing Full Well

Ernest Sosa

"A truly excellent book that will be widely praised across contemporary philosophy. . . . Sosa is without doubt the most influential figure in modern epistemology, and this book will be required reading for anyone who wants to know the current state of play in the field."
—Duncan Pritchard, University of Edinburgh

Soochow University Lectures in Philosophy
Cloth $29.95 978-0-691-14397-2

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 3
MARCH 2011

page
111 *Skepticism about Moral Expertise as a Puzzle for Moral Realism* Sarah McGrath
138 *How to Be a Pessimist about Aesthetic Testimony* Robert Hopkins
158 *Separating Directives and Assertions Using Simple Signaling Games* Kevin J. S. Zollman

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2011 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 3, MARCH 2011

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA

SKEPTICISM ABOUT MORAL EXPERTISE AS A PUZZLE FOR MORAL REALISM*

In this paper, I explore a neglected puzzle for the moral realist. I then canvass some potential responses. Although I endorse one response as the most promising of those I survey, my primary goal is to make vivid how formidable the puzzle is, as opposed to offering a definitive solution.

I. MORAL DEFERENCE

There is something off-putting about the idea of arriving at one's moral views by simply deferring to an expert. By contrast, there is no problem with deferring to a tax specialist about one's taxes. In the case of taxes, if one wants to bypass the struggle of figuring out what one owes, one can simply pile all of the relevant information into a box and drop it off for someone else to process. But there seems to be something problematic about bypassing the struggle of figuring out what one owes, morally speaking, and leaving this for someone else to determine. The asymmetry can appear puzzling. If it is fine to defer to a tax specialist when one's finances get complicated, why is it not equally appropriate to defer to a moral specialist when one's life gets complicated in morally significant ways?

Of course, the extent to which moral deference is appropriate is a controversial matter. Disagreement about this general issue is among the key theological divisions between Catholicism and Protestantism, and this disagreement is echoed in philosophy. The Protestant tradition, in prioritizing the individual conscience over the dictates of

---

*Earlier versions of this paper were presented at the University of Notre Dame, Temple University, Princeton University, and at the Eastern Division meetings of the American Philosophical Association; I am grateful to those audiences for their feedback. In addition, I would like to thank Robert Audi, Elizabeth Harman, Thomas Kelly, and the anonymous referees for this JOURNAL for helpful suggestions.

0022-362X/11/0803/111-137

would-be mediating authorities, finds perhaps its clearest secular reflection in the Kantian emphasis on the moral significance of *autonomy*. Thus, writing within the broadly Kantian tradition, Robert Paul Wolff commends the practice of the "autonomous man," which he describes as follows:

> The autonomous man...may learn from others about his moral obligations, but only in the sense that a mathematician learns from other mathematicians—namely by hearing from them arguments whose validity he recognizes even though he did not think of them himself....He does not learn as one learns from an explorer, by accepting as true...accounts of things he cannot see for himself.[1]

By contrast, writing within the broadly Catholic tradition, Elizabeth Anscombe defends the view that there is nothing problematic about deference to a genuine moral authority—in the form of church teaching, or otherwise:

> A man may have reason to judge that another man's moral counsel is more reliable than his own unaided conscience...he may, moreover, have reason to believe that some public source of moral teaching is more reliable than his own judgment.[2]

In such cases, Anscombe suggests, the reasonable course is to defer to the source that one takes to be more reliable. Others take the same view: although undoubtedly a strange bedfellow for Anscombe in some respects, Peter Singer has similarly defended the idea that there is nothing especially problematic about moral expertise, and thus, presumably, about moral deference.[3]

There are, then, deep disagreements about the place of expertise within the moral domain. Significantly, however, even many of those who are most sympathetic to the idea of genuine moral expertise assume that an important aspect of the expert's role is that of providing reasons why, for example, a given type of conduct is morally wrong.[4] The thought is that ideally, those to whom the teaching is addressed would recognize these reasons as compelling, and as providing a sufficient basis for refraining from the conduct in question. Where such recognition is not forthcoming, things are less than ideal. By contrast, it does not seem particularly important that one have much of a grasp on how exactly the tax specialist arrives at his estimate of what one owes: even if one treats his role in the process as essentially a black-box mechanism, nothing of significance is lost.

---

[1] Robert Paul Wolff, *In Defense of Anarchism* (New York: Harper and Row, 1970), p. 13.

[2] G. E. M. Anscombe, "Authority in Morals," in *The Collected Philosophical Papers of G. E. M. Anscombe Volume III: Ethics, Religion, and Politics* (Oxford: Blackwell, 1981), pp. 46–47.

[3] See especially Peter Singer, "Moral Experts," *Analysis*, xxxii, 4 (March 1972): 115–17, and "Ethical Experts in a Democracy," in David Rosenthal and Fadlou Sehadi, eds., *Applied Ethics and Ethical Theory* (Salt Lake City: Utah UP, 1988), pp. 149–61.

[4] This assumption seems clearly reflected in, for example, papal encyclicals, as well as in the practice of the kind of secular expert ethics panels that Singer is concerned to defend. As C. A. J. Coady (another philosopher broadly sympathetic to the possibility of acquiring moral knowledge via testimony) notes, "When someone tells me that what I am proposing to do is immoral, I do not react by asking for his credentials but for his reasons." Coady, *Testimony* (New York: Oxford, 1994), p. 71.

Let us reserve the term *deference* for cases in which one holds a view solely because another person holds that view. For example, if I believe, solely on the basis your testimony, that we should turn left rather than right at the next light in order to reach our destination, then I defer to you about which way we should turn.[5] Similarly, suppose that I believe that capital punishment is wrong, not because of a conviction that it has such-and-such features which render it morally impermissible, but rather simply because you told me that it is wrong. (Perhaps if left to my own unaided judgment, I would believe that capital punishment is permissible, or else have no opinion at all about its moral status, in the same way that I might have no opinion about which way we should turn in the absence of your direction.) In that case, I defer to you as to the morality of capital punishment.

Again, a potentially puzzling asymmetry looms. At the very least, there seems to be something odd or peculiar about my genuinely believing that capital punishment is wrong simply on the basis of your testimony, in a way that there is nothing similarly odd or peculiar about my believing that we should turn left simply on the basis of your testimony, or my believing that I owe a certain amount of money to the government on the basis of my accountant's testimony. Of course, there are at least some cases of moral deference that everyone will agree are unproblematic. For example, I might defer to your judgment on some moral question if I know that you possess relevant nonmoral information that I lack. Our moral views about which Nazis committed the most egregious atrocities are sensitive to the historical information we possess about who was responsible for giving which orders, and so on. If I know that we possess similar moral sensibilities but that you are better informed about the relevant historical facts, then it makes sense for me to defer to your judgments about the relative moral turpitude of individual Nazis. In this kind of case, although I defer to your moral judgments, I do not treat you as a *moral* expert in the relevant sense. Your relative expertise concerns history and not morality: insofar as there is a sense in which you do possess superior moral expertise in virtue of your superior historical knowledge, it is expertise of a highly derivative sort.[6]

---

[5] Notice that here, holding a view "solely because another person holds that view" does not imply or suggest that one lacks independent reason to think that the person is reliable. (Indeed, my decision to defer to you about which way we should turn might be due to my background knowledge that your sense of direction is superior to mine.)

Perhaps there are other cases in which moral deference is as straightforward as nonmoral deference. Suppose that I know that you and I share the same moral sensibility, but that with respect to some particular question that is before us my judgment is impaired in a way that yours is not. Perhaps I am too angry or too drunk to think clearly, or I am too personally involved to see things aright. In such circumstances, I might view deference to your judgment as deference to my unimpaired self, in much the way that I view deference to your historically informed judgments about the relative turpitude of particular Nazis as deference to my better-informed self. Similarly, two people who disagree about what constitutes a fair division of goods between them might ask a passing stranger to settle the dispute, not because they suspect that the stranger possesses some moral wisdom or knowledge that they lack, but rather because the stranger's judgment is relatively unbiased.[7]

Let us call cases such as these *impure* cases of moral deference. In contrast, a pure case of moral deference would be a case of the following sort:

> I possess all of the nonmoral information thought relevant to the moral permissibility of the death penalty that you possess, and I have no special reason to think that my judgment is impaired. On the basis of this information and my own careful consideration of the issue, either it seems to me that the death penalty is morally permissible, or I remain undecided. Nevertheless, I judge that the practice is not permissible, on the basis of your testimony to that effect.

---

[6] Cf. Allan Gibbard's discussion of "contextual" authority, which "presupposes a context of shared norms." He writes:

> Might I take someone else's accepting something normative as my own reason for accepting it? In some cases I might well. Suppose that you tell me it made no sense for Cleopatra to be angry at the messenger. I am ignorant of history, perhaps, and confident that you know your history, and that you and I share the same basic norms for anger. In that case, I can take your normative reasoning as proxy for my own. I think that you are reasoning just as I would if I knew the facts. I can let you draw normative conclusions for me, and so I take the fact that you draw a normative conclusion as reason for accepting it myself.

Gibbard, *Wise Choices, Apt Feelings* (Cambridge: Harvard, 1990), p. 174. As this passage suggests, Gibbard is not specifically concerned with the case of morality but with the normative realm more generally.

[7] The last example is borrowed from Montaigne via Julia Driver, "Autonomy and the Asymmetry Problem for Moral Expertise," *Philosophical Studies*, cxxviii, 3 (April 2006): 619–44.

In such a case, I in effect treat you as possessing purely moral information that I lack. In this respect, pure moral deference involves treating someone as a genuine moral expert in a very strong sense.[8]

Consider then

> The datum: Pure moral deference seems more problematic than deference in many other domains.

In what follows, I will explore what the datum suggests about the nature of morality. As we shall see, certain meta-ethical views offer ready explanations for the datum, while others struggle to account for it. In particular, for reasons that I offer in the next section, the datum is at least initially puzzling given a broadly realist view about morality, in a way that it is not at all puzzling on various rival views. Before we consider its potential significance, however, a few clarificatory remarks about the datum are in order.

First, notice that the datum makes no mention of moral testimony *per se*. Although it is sometimes claimed that there is something strange about moral testimony, on the present view any such strangeness is a special case of a more general phenomenon.[9] While testimony is a particularly common and salient route by which we learn the moral views of others, there are, of course, other routes. Suppose that you actively attempt to conceal your opposition to capital punishment from me, perhaps out of an overzealous desire not to interfere with the exercise of my own autonomous moral judgment. Nevertheless, I discover that you strongly disapprove of capital punishment—perhaps by observing your involuntary grimaces whenever others mention the practice in your presence. As a result, I adopt your strongly negative attitude as my own. On the face of it, this seems no less (and no more) odd than a case in which I defer to you about the morality of capital punishment after learning of your view on the basis of testimony.

Second, notice that the datum seems to hold for moral judgments of varying levels of generality and of various kinds. On the one hand, it seems that there is something peculiar about my believing the general moral principle that capital punishment is wrong if the only basis for my conviction is the fact that this is what you believe. But it seems no less peculiar for me to defer to your judgment about the moral status of some token action or event, for example, the wrongness of a particular execution. Similarly, it seems that there would be something odd about my adopting a belief about what I myself have most moral reason to do on the basis of pure moral deference. For example, it would be extremely odd for me to believe that I have strong moral reasons to avoid eating meat solely on the basis of your say-so, where this is not a matter of my suspecting that you are better informed about the relevant nonmoral facts, or that I am in the grip of some kind of bias. But as some of the other examples that we have considered make clear, the same sense of peculiarity seems to attach to deference regarding moral judgments that do not directly concern one's own behavior. In these respects, the scope of the phenomenon seems quite broad.

---

[8] The distinction between impure and pure moral deference corresponds, at least roughly, to Gibbard's distinction between "contextual authority" (see note 6) and "fundamental authority" as discussed in *Wise Choices, Apt Feelings*, pp. 174–81 According to Gibbard, "even if it is limited and mutual, fundamental authority is puzzling" in a way that contextual authority is not (p. 176).

[9] See, for example, Philip Nickel, "Moral Testimony and Its Authority," *Ethical Theory and Moral Practice*, IV, 3 (September 2001)· 253–66, and Robert Hopkins, "What Is Wrong with Moral Testimony?" *Philosophy and Phenomenological Research*, LXXIV, 3 (2007): 611–34. For defenses of moral testimony, see Karen Jones, "Second-Hand Moral Knowledge," this JOURNAL, XCVI, 2 (February 1999): 55–78; and Coady, *op. cit.*, pp. 69–75.

In another respect, however, the datum is relatively weak: the claim is not that pure moral deference is more problematic than nonmoral deference, but rather that it *seems* more problematic. Perhaps the seeming strangeness of moral deference is mere appearance, and deference about morality is (or ought to be) as straightforward a matter as deference about local geography. That is, perhaps our sense that there is something odd or peculiar about moral deference is simply mistaken. Even if that is the case, it is still true that moral deference strikes us, or many of us, as peculiar in a way that deference about geography does not. Given this, we can ask what explains this fact.

II. COGNITIVIST AND NONCOGNITIVIST EXPLANATIONS OF THE DATUM

Cognitivism is the view that moral sentences such as "capital punishment is wrong" typically express truth-evaluable propositions when uttered on particular occasions; in sincerely uttering the sentence, one expresses a belief that one holds. Noncognitivists deny these claims. According to the noncognitivist, when one utters the sentence, "Capital punishment is wrong," one is not describing anything or saying something that is true or false. Thus, according to the brand of emotivism associated with Ayer, one is expressing one's disapproval of capital punishment; according to Stevenson, one is expressing a kind of "unfavorable interest" in capital punishment; according to Gibbard, one is expressing one's acceptance of a system of norms that proscribe the practice.[10] Notice that noncognitivist views seem to provide ready explanations of the datum. For example, suppose that my judging that capital punishment is wrong is, as Ayer would have it, a matter of my expressing my own negative emotions towards capital punishment. In that case, it is unsurprising that we find something extremely odd about the idea of my deferring to you about whether capital punishment is wrong.[11]

---

[10] See A. J. Ayer, *Language, Truth, and Logic* (London: Victor Gollancz, 1936); Charles L. Stevenson, *Ethics and Language* (New Haven: Yale, 1944); and Gibbard, *op. cit.*

Similarly, the *constructivist* would seem to be in a strong position to account for the datum. For the constructivist, like the noncognitivist, will insist that it is a deep confusion to suppose that there is some independent realm of moral facts or truths that we are attempting to bring into view and to which our moral judgments are answerable. For example, a Kantian constructivist might hold that, whenever we engage in pure moral deference, we are guilty of misconstruing the nature of the moral domain: we are in effect treating the moral domain as a repository of antecedently existing facts, as opposed to something that we must construct by the autonomous exercise of our own reason.[12]

On the other hand, the datum seems to pose at least a *prima facie* challenge for certain forms of cognitivism. After all, when it comes to deep and unobvious facts about the empirical world, we readily defer to others (in some cases scientists, in other cases explorers, and so on) who are better placed to discover those facts than we are. In such cases, even a very sweeping kind of deference to expert opinion seems appropriate. If, similarly, there is a domain of deep and unobvious moral facts, then it is natural to expect that some of us—intuitively, the "moral experts"—would be better placed to discover these facts than others. In principle then, moral deference should strike us as no more peculiar than deference about scientific matters or geography.[13]

---

[11] Cf. David McNaughton, *Moral Vision* (Malden, MA: Blackwell, 1988), p. 10 It is an interesting question, although not one that I will explore here, whether broadly noncognitivist views that seek to appropriate the realist's talk of moral truth and related notions retain this apparent virtue. See, for example, Simon Blackburn, *Essays in Quasi-Realism* (New York: Oxford, 1993)

[12] For clear statements of the constructivist rejection of the idea that the aim of moral judgment is to depict antecedently existing moral facts, see, for example, Christine Korsgaard, "Realism and Constructivism in Twentieth-Century Moral Philosophy," in *The Constitution of Agency* (New York: Oxford, 2008), pp. 302–26; John Rawls, "Kantian Constructivism in Moral Theory," reprinted in his *Collected Papers* (Cambridge: Harvard, 1999), pp. 303–58; and Sharon Street, "What Is Constructivism in Ethics and Metaethics?" *Philosophy Compass*, v, 5 (May 2010): 363–84.

[13] Could the cognitivist claim that it is the *a priori* character of morality that explains our ambivalence towards the possibility of full-fledged moral expertise? According to this line of thought, if the most fundamental moral truths are knowable independently

Of course, there are domains in which a commitment to cognitivism coexists easily with the sense that there would be something very odd about deferring to another person. Consider, for example, domains in which the truth is readily available to anyone, or almost anyone. Thus, it would be odd to defer to someone else about simple claims of arithmetic, but this does not suggest that such claims are neither true nor false. Rather, the reason why deference about simple arithmetical truths seems out of place is the apparent lack of need for such deference. Could the moral cognitivist offer an explanation along these lines?

In fact, a venerable tradition in moral philosophy insists that everyone (or at least every nondefective adult human being) is in a position to know the essential truths of morality, independently of the assistance of others.[14] Here, for example, is Thomas Reid:

> From these self-evident first principles the whole system of moral conduct follows so easily, and with so little aid of reasoning, that every man of common understanding, who wishes to know his duty, may know it. The path of duty is a plain path, which the upright in heart can rarely mistake. Such it must be, since every man is bound to walk in it. There are some intricate cases in morals which admit of disputation; but these seldom occur in practice; and when they do, the learned disputant has no great advantage....In order to know what is right and wrong in human conduct, we need only listen to the dictates of our conscience, when the mind is calm and unruffled, or attend to the judgment that we form of others in like circumstances.[15]

The general idea that the moral facts are in some sense available to everyone has also been endorsed by a number of contemporary

---

of experience, then morality will differ from empirical domains in the extent to which experience can render some better placed than others to discover those truths. But on the face of it this strategy seems unpromising: after all, we are certainly not wary about experts in the apparently *a priori* domain of mathematics. Believing Fermat's last theorem on the basis of deference to an expert mathematician seems on a par with believing the theory of special relativity on the basis of deference to the physicist. I discuss this issue more extensively in "The Puzzle of Pure Moral Deference," in John Hawthorne, ed., *Philosophical Perspectives 23: Ethics* (Malden, MA: Wiley, 2009), pp. 321–44

[14] For references and further discussion of this tradition, see *ibid.* A main theme of J. B. Schneewind's *The Invention of Autonomy* (New York: Cambridge, 1988) is the extent to which the development of modern moral philosophy was driven by the desire for an account of morality that would make an awareness of its requirements universally available.

[15] Thomas Reid, *Essays on the Active Powers of the Human Mind.* First published 1788; excerpt reprinted in Schneewind, ed., *Moral Philosophy from Montaigne to Kant* (New York: Cambridge, 2003), pp. 632–50, at p. 645.

philosophers, including Bernard Gert,[16] Nicholas Rescher,[17] and Michael Smith.[18]

Could the cognitivist appeal to the alleged universal availability of the truths of morality in order to account for the datum? Elsewhere, I have argued that this strategy founders on a dilemma.[19] In brief, the dilemma runs as follows. Consider two different interpretations of the claim that the truths of morality are universally available. On the first interpretation, the truths of morality are, like the simplest truths of arithmetic, literally trivial: they are obvious or potentially obvious to anyone who is genuinely concerned to know them. If the truths of morality were universally available in this sense, then the cognitivist would be in a position to account for the datum in a satisfying way. However, such a characterization of our cognitive relationship to morality is implausibly optimistic. Notably, the diversity of opinion that we find with respect to many moral issues belies the idea that the truths of morality are obvious or even potentially obvious; *pace* Reid, genuine moral disagreement is neither an uncommon nor a marginal phenomenon. Moreover, there seems to be no good reason to deny that, among those who disagree about issues such as the morality of abortion or capital punishment, at least some on each side of the debate are people of good will who sincerely desire to know what morality requires. Even apart from interpersonal considerations, genuine moral uncertainty seems far more common than would be the case on the suggested picture.

Consider an alternative interpretation of the thought that the truths of morality are universally available. On this interpretation, even though some moral truths are unobvious, those truths are nonetheless in principle accessible to all normal adult human beings, in the following sense: none of us lacks the *wherewithal* to discern those truths, in a way that not all of us have the wherewithal to discern the truths of quantum electrodynamics or algebraic topology. On this view, all normal adult human beings have a level of conceptual sophistication and general intelligence that is sufficient to autonomously grasp the truths of morality, provided that that intelligence and sophistication is properly employed. A proponent of this view might maintain that the bulk of moral disagreement is due to certain common biases, false nonmoral beliefs, and simple performance errors that distort our moral reasoning, and that our moral views would converge under suitably idealized conditions of reflection and deliberation.

---

[16] "Morality does not require beliefs that are not known to all moral agents." Gert, *Common Morality* (New York: Oxford, 2004), p. 90. Cf. p. 3, where Gert asserts that "this book. .contains no new information about what kinds of actions morality prohibits, requires, discourages, encourages, or allows. Anyone who is intelligent enough to read this book already has all of this information."

[17] "Anything that requires extensive knowledge or deep cogitation is *ipso facto* ruled out as a moral precept or principle. The very idea of moral expertise…is for this reason totally unrealistic." Rescher, *Commonsense* (Milwaukee, WI: Marquette, 2005), p. 200.

[18] "Our moral life seems to presuppose that [moral] facts are in principle available to all; that no one in particular is better placed to discover them than anyone else." Smith, *The Moral Problem* (Malden, MA: Blackwell, 1994), p. 5.

[19] McGrath, *op. cit.* The discussion of the next three paragraphs draws on material from section III of that paper.

This view about morality is at least somewhat more plausible than the view according to which morality consists of a body of trivialities. However, notice that even if it were true, it would not provide a good explanation of the datum. For it often makes perfect sense to defer to someone, even if one has the capacity or wherewithal to figure out the answer for oneself. For example, I might defer to my accountant with respect to the question of how much money I owe the government, even if I am perfectly capable of figuring out the answer on my own given sufficient attention to the task. In contrast to pure moral deference, this seems like a paradigmatic case in which deference to another person is completely straightforward and unproblematic. Hence, the dilemma for such "universalist" approaches to explaining the datum: the more a given view makes the moral facts obvious to all, the greater its ability to explain the datum, but the less plausible it is; on the other hand, the more a view allows for the unobviousness of moral facts by idealizing the sense in which those facts are universally available, the more plausible it becomes, but the less it is able to function as an explanation.

When the truth is obvious, deference seems out of place because it is otiose. At the other end of the spectrum, truths which none of us are in a position to know similarly generate contexts in which deference is out of the question, without thereby creating any presumption against a cognitivist interpretation of the relevant domain. Even if claims about the outcomes of future coin tosses and lotteries are true or false, we do not believe that there are experts about such matters, because we do not believe that anyone is in a position to know whether the claims are true. But again, an explanation along these lines does not seem to hold much promise in the moral case. For it is not only moral skeptics, those who deny that we are in a position to attain genuine moral knowledge, who find something odd about the idea of pure moral deference.

Consider next a more philosophically interesting domain in which a commitment to cognitivism coexists with a sense that deference to others is typically out of place: the domain of self-knowledge. It is a familiar idea that we enjoy a kind of privileged access to our own mental states. Indeed, a historically popular claim is that one's present-tense beliefs about one's own mental states are incorrigible: according to this line of thought, if I believe that I am in pain on the basis of introspection, this belief is not subject to correction by others. Even if this incorrigibility thesis is too strong, we can certainly agree that, at the very least, it would be extremely odd for me to defer to someone else about whether I am in pain right now, or (more generally) to treat that person as possessing superior expertise about what mental states I am in. Yet this does not lead us to think that utterances about one's own mental states do not express propositions.[20] Thus, the domain of self-knowledge is one in which a commitment to cognitivism coexists easily with the sense that there would be something very odd about treating others as experts.[21]

How useful should the cognitivist find the example of self-knowledge? Some cognitivists might find the example extremely useful. Consider, for example, a toy subjectivist or individual relativist theory, according to which an utterance of the sentence, "Capital punishment is wrong," expresses the truth-evaluable proposition that the speaker disapproves of capital punishment. Notice that such a view readily explains the seeming oddity of pure moral deference. On the face of it, each particular agent is typically best placed to find out whether an utterance of "capital punishment is wrong" in his or her own mouth is true, because he or she is typically best placed find out whether he or she disapproves of capital punishment at the time of the utterance. Indeed, on this view, the oddity of pure moral deference turns out to be a special case of the oddity of deferring to someone else about one's own mental states.

---

[20] Here I simply assume the orthodox view of the matter, while noting that it has been contested. Thus, according to a view sometimes associated with Wittgenstein's *Philosophical Investigations* (Oxford: Basil Blackwell, 1953), apparent self-ascriptions of pain are not actually in the business of stating facts; rather, in saying "I'm in pain," one is expressing one's pain, much as if one had said "Ouch!" instead. One of the reasons for the popularity this view once enjoyed seems to have been a commitment to (something like) the incorrigibility thesis, along with the thought that, if utterances of the relevant kind were in the business of stating facts, their incorrigibility would be inexplicable. This line of thought—that a nonfactualist construal of the relevant kind of discourse is the best explanation of its peculiar status—is parallel to the one at issue in this paper.

[21] Of course, there are exceptional contexts in which a person might treat someone else as an expert about at least some of her mental states: for example, a person might very well defer to her therapist about whether she is actually in certain mental states (although presumably not whether she is currently experiencing physical pain). Notwithstanding such exceptional contexts, it is clear that deference about one's own mental states is a relatively marginal phenomenon compared to the kind of sweeping deference that we find in many other domains, and that a natural explanation of this is the fact that one is typically better positioned than others to make judgments of the relevant kind.

Thus, at least some possible cognitivist views accommodate the datum as readily as noncognitivism or constructivism. The same does not seem to be true, however, of moral realism. Providing a fully satisfactory characterization of moral realism is a subtle matter; for our purposes, we can make do with the following (admittedly rudimentary) gloss.[22] First, the moral realist is a cognitivist, in that she holds that moral sentences express truth-evaluable propositions. But in contrast to the subjectivist or individual relativist, the realist holds that the moral facts are objective, in that the moral status of a practice such as capital punishment is independent of our moral opinions, or which moral framework we accept.[23] Moreover, although the moral realist will readily admit that at least some moral facts are unobvious, she is not a moral skeptic: she does not hold that our best efforts to achieve genuine moral knowledge are doomed to failure. Rather, the moral realist holds that although the discovery of some moral facts might be an intellectually demanding task, we are capable of attaining, and in fact possess, substantial moral knowledge.[24]

---

[22] For some of the subtleties, see Smith, "Moral Realism," in Hugh LaFollete, ed., *The Blackwell Guide to Ethical Theory* (Malden, MA: Blackwell, 2000), pp. 15–37; Russ Shafer-Landau, *Moral Realism* (New York: Oxford, 2003), especially pp 13–18; and Geoffrey Sayre-McCord, "Moral Realism," in David Copp, ed., *The Oxford Handbook of Moral Theory* (New York: Oxford, 2006).

[23] Thus, according to Shafer-Landau, "realism ... insists that the truth of any first-order normative standard is not a function of what anyone happens to think of it." Shafer-Landau, *op. cit.*, p. 15. Similarly, Richard Boyd characterizes realism partially in terms of whether the truth of a moral belief is independent of our moral opinions and theories in his "How to Be a Moral Realist," Sayre-McCord, ed , *Essays on Moral Realism* (Ithaca· Cornell, 1988), p. 182. And J. L. Mackie, in characterizing what he takes to be the ordinary person's commitment to realism, claims that on this view, when a person makes a moral claim he says "something about whatever it is that he characterizes morally, for example, a possible action, as it is in itself, or would be if it were realized, and not about, or even simply expressive of, his, or anyone else's, attitude or relation to it." Mackie, *Ethics Inventing Right and Wrong* (New York. Penguin, 1977), p. 33.

[24] In "Objectivity and Modern Idealism: What Is the Question?" Michaelis Michael and John O'Leary-Hawthorne, eds., *Philosophy in Mind: The Place of Philosophy in the Study of Mind* (Boston: Kluwer, 1994), pp. 277–319, Gideon Rosen contends that we lack any satisfactory way of giving content to the (alleged) distinction between realism and anti-realism about a given domain. While he readily admits that there is a clear distinction between being a "minimal realist" (that is, someone who holds that "our core beliefs about the disputed subject matter are more or less true," p. 281) and being a noncognitivist or error theorist, he is skeptical that there is some good way of further dividing the camp of minimal realists into realists and anti-realists in terms of mind-dependence, objectivity, and so on. It is possible that my "moral realist" is simply a minimal realist about morality in Rosen's sense, given the understanding that the different versions of subjectivism and relativism are excluded. In that case, the extension of "moral realist" as it used in this paper would be relatively inclusive. Still, the puzzle in question would not be *everyone's* problem· at least some noncognitivists, for example, would have an immediate and obvious solution to it.

Notice that the moral domain as the realist conceives of it seems to be exactly the kind of domain where one would expect to find different levels of expertise among different individuals.[25] On the realist's view, there are moral facts, which a given individual might either know or not know. In general, these moral facts are neither trivial nor completely unknowable. Further, the objective character that moral facts are alleged to have undermines the possibility that each individual is his or her own best moral authority, as would be the case if some variety of subjectivism or individual relativism were true. In these respects, then, the moral domain as the realist understands it seems to be one in which some individuals would count as experts relative to others. We thus might expect to find pure moral deference as straightforward a practice as deference anywhere else. The fact that we do not find this is at least initially surprising, given a realist view. In short, the datum seems to cast at least some *prima facie* doubt on the general picture of morality that the realist presents.

How should the realist reply? Let us consider some possibilities.

III. DOES MORAL KNOWLEDGE REQUIRE MORAL UNDERSTANDING?

Suppose that Andrew and Benjamin are mid-nineteenth-century abolitionists who are both passionately opposed to slavery. However, while Benjamin's opposition is based on his own sensitivity to the reasons why slavery is wrong, Andrew's opposition is based solely on the testimony of someone whom he takes to be a reliable moral authority. What account might the realist offer for the sense that there is something extremely odd about Andrew's stance with respect to slavery compared to Benjamin's?

According to one class of answers that the realist might offer, the difference resides in the inferior epistemic status of Andrew's belief that slavery is wrong. For example, the realist might suggest that one can only *know* that slavery is wrong if one's belief is at some level based on those considerations in virtue of which slavery is wrong: in order to genuinely know that slavery is wrong, one must have at least some appreciation of the reasons why it is. More generally, the realist might suggest that *moral knowledge requires moral understanding*. Notice that, in general, when one holds a true belief solely on the basis of another's testimony, one does not believe the relevant proposition on the basis of those considerations that underwrite its truth: if I believe that *p* is true solely on the basis of your say-so, I typically do not understand why *p* is true. Thus, pure moral deference does not deliver moral understanding. If moral understanding is required for moral knowledge, pure moral deference is not a possible source of moral knowledge: while Benjamin's belief is a genuine candidate for knowledge, Andrew's belief is not.

---

[25] Cf. David Brink, *Moral Realism and the Foundations of Ethics* (New York: Cambridge, 1989), p. 95. Brink briefly considers the possibility that someone might reject moral realism on the grounds that it has "the implausible implication that there are moral experts or authorities"; in response, he offers a qualified defense of moral expertise, and thus attempts to turn the point to the moral realist's advantage (pp. 95–98).

Consider the analogous view about mathematics. Notably, both Roderick Chisholm[26] and Bernard Williams[27] have defended the view that one cannot have full-fledged mathematical knowledge on the basis of another's testimony. Someone who believes that a given mathematical proposition is true on the basis of grasping a proof of it has genuine mathematical knowledge in a way that someone who believes the same proposition on the basis of testimony does not. In the mathematical case, there is a particularly intimate relationship between the relevant mathematical truth and a canonical proof of it. Similarly, there seems to be a particularly intimate relationship between the fact that slavery is wrong and those features of slavery in virtue of which it counts as wrong. By contrast, whatever justification is afforded by testimony in the two cases seems extremely extrinsic to the truth of the proposition in question. According to the current proposal, the acquisition of genuine moral knowledge requires one to engage in moral deliberation, or at least in rational processes that, even if more rudimentary than deliberation, nevertheless resemble it in that they too involve sensitivity to those considerations that underwrite the truth of the relevant proposition. On this view, our attitude about moral deference stems from our perhaps tacit recognition that such deference is not a genuine source of moral knowledge.

However, on balance this proposal is not an attractive option for the moral realist. First, notice that, at least outside of the moral domain, one often knows that *p* is the case even if one has no understanding as to why *p* is the case. (One can know, even know with certainty, that drinking saltwater causes dehydration, even if one has no idea why drinking saltwater causes dehydration.) Indeed, situations with this structure often set the agenda for substantive inquiry in the sciences and elsewhere. In general, it is simply not true that one's reasons for believing *p* must be among the reasons why *p* is true in order for one's belief to count as knowledge. That is why testimony can deliver knowledge of empirical matters even in the absence of any explanation as to why things are as the testifier describes them. Why then should morality be different? Perhaps some reason might be offered for thinking that moral knowledge differs from empirical knowledge in this respect. But in any case, it is important to see that any such reason should not have much appeal for the *realist*. For the realist, that a particular token action has the property of being wrong is an objective fact about it. If one knew that someone else was a reliable detector of that property, then surely one could gain knowledge that the action possessed the property on the basis of her testimony, just as one could gain knowledge that some difficult to detect natural property is instantiated in such-and-such circumstances from the reliable testimony of a natural scientist.

---

[26] Chisholm, *Theory of Knowledge* (Englewood Cliffs, NJ: Prentice Hall, 1966).

[27] Williams, "Knowledge and Reasons," in G. H. von Wright, ed., *Problems in the Theory of Knowledge* (The Hague, Netherlands: Nijhoff, 1972), pp. 1–11 Ayer also seems to implicitly accept the view in his "Comments on Williams," at pages 12–16 of the same volume. For criticism, see especially Coady, "Mathematical Knowledge and Reliable Authority," *op. cit.*, pp 249–61.

Thus, the realist should reject this strategy: she should allow that pure moral deference can deliver moral knowledge, even though it does not typically deliver moral understanding. More generally, the realist should reject the idea that the difference between Andrew and Benjamin resides in the epistemic status of their respective beliefs.[28]

A similar point holds for more radical proposals, for example, that a moral view cannot be fully understood if it is held on the basis of another's testimony. This is the view of Philip Nickel:

> This is why many beliefs gained from moral dependence are deficient: such claims, if relied on without independent justification, are not fully understood.[29]

Elsewhere, Nickel suggests that full understanding requires "a sensitivity to justification conditions" and that this is what moral testimony often fails to deliver.[30] As an example of a moral claim that is not fully understood when it is believed on the basis of testimony, he offers

---

[28] Nor should the realist say that the datum is explained by the fact that moral knowledge based on understanding is epistemically *better* knowledge than moral knowledge where such understanding is lacking. To the extent that there are ways in which moral knowledge plus understanding is better than moral knowledge alone, the same seems true of nonmoral domains.

The suggestion that moral knowledge requires moral understanding is also criticized by Alison Hills, "Moral Testimony and Moral Epistemology," *Ethics*, cxx, 1 (October 2009): 94–127.

[29] Nickel, *op. cit.*, pp. 260–61.

[30] *Ibid.*, p. 259. Notice that there seem to be two ideas in play here (i) in order to fully understand a moral claim one must actually possess some independent justification for that claim, and (ii) in order to fully understand a moral claim one must be sensitive to its justification conditions; that is, one must possess the ability to recognize a justification for the claim, if such a justification were encountered. The second requirement is significantly weaker than the first. But the point made in the main text holds against either interpretation. Nickel's ultimate conclusion is that *actions* undertaken in response to moral testimony cannot be "morally good." For discussion of proposals of this general sort, see section v below.

"Socrates should have fled prison." On the face of it, the suggestion that anyone who holds this belief on the basis of testimony must fail to fully understand it is an implausible one. But whatever arguments might be offered on its behalf, the *realist* should reject the relevant picture of understanding out of hand. In general, the idea that one can understand a claim only if one independently grasps its justification conditions is a characteristic thesis of anti-realism; indeed, disagreement on this point is sometimes seen as definitive of the dispute between realist and anti-realist.[31] Thus, the realist should reject this picture of what is required for understanding a moral claim, on pain of compromising his realism beyond recognition.

IV. RECOGNIZING MORAL EXPERTS

Let us turn to an alternative realist strategy, according to which moral deference is, at least in principle, no more problematic than deference in any other domain. On this view, the datum is explained by the fact that, in practice, formidable epistemological difficulties arise when one attempts to identify genuine moral experts as such. Our sense that there is something odd about pure moral deference is an artifact of our (perhaps implicit) recognition that, in the moral case, one typically lacks the kinds of compelling grounds for attributing genuine expertise that one often possesses in other domains.[32]

More specifically, notice that it is often easier to acquire clear, unequivocal evidence that someone possesses reliable judgment in various empirical domains than it is to acquire evidence that someone possesses reliable judgment in the moral domain. Consider, for example, weather forecasting. Even if one knows nothing about weather-forecasting techniques and is wholly ignorant of meteorology, it is a relatively straightforward matter to collect evidence that bears directly on the reliability of rival weather forecasters. When I observe that it is raining today, I acquire a piece of evidence that speaks in favor of the reliability of those forecasters who predicted rain and against the reliability of those who failed to do so. In this way, I can amass clear track-record evidence relevant to the question of who possesses genuine weather-forecasting expertise.

---

[31] See, for example, Michael Dummett, *The Logical Basis of Metaphysics* (Cambridge: Harvard, 1991).

[32] The idea that there are deep difficulties in identifying genuine moral experts as such (even on the assumption that such experts exist) is an ancient one; it is clearly present in Plato's early dialogues. For useful discussion, see Alexander Nehamas, *The Art of Living* (Berkeley: California UP), especially pp 78–82, and "Socratic Intellectualism," reprinted in his *Virtues of Authenticity* (Princeton: University Press, 1999).

When one observes that it is raining, one in effect has independent access, via perception, to facts in the target domain. In arriving at the view that a given weather forecaster correctly predicted rain, one does not rely on techniques similar to those employed in arriving at that prediction. Because one is in a position to determine reliably which weather forecasters got it right on particular occasions by some method other than those employed by the weather forecasters themselves, one has a means of calibrating their techniques for accuracy. In a similar way, a scientist might calibrate a new instrument for accuracy by checking the reliability of its verdicts about some range of facts to which he has independent access.[33] By contrast, there seems to be no analogous way to calibrate the accuracy or reliability of someone's moral judgment, because one lacks the relevant kind of independent access to the moral facts. If one attempted to rank others with respect to the accuracy of their moral judgment by checking how often they answered controversial moral questions correctly, it seems as though one could do so only by engaging in first-order moral reasoning and deliberation of one's own. It is thus unsurprising that clear and unequivocal evidence that someone possesses unusually reliable moral judgment is hard to come by. Even if there are genuine moral experts, locating particular individuals within the space of moral expertise is undoubtedly a precarious business.[34]

Of course, insofar as one holds opinions about disputed moral questions, one will also hold that some individuals have more reliable moral judgment than others, on the basis of their having arrived at (what one takes to be) the correct answers to those questions. But there is an important disanalogy between the weather-forecasting case and the moral case. The disanalogy is not that one's assessment of who possesses reliable moral judgment depends on one's own first-order moral views. That much is true, but it does not distinguish the two cases: by the same token, one's beliefs about who is a reliable weather forecaster are not independent of one's own first-order beliefs about the weather. (If one did not believe that it is raining today, one would not count this fact as being to the credit of those weather forecasters who predicted rain and to the discredit of those who failed to do so.) Rather, the difference seems to be that we lack independent access to the moral facts, and so cannot accumulate track-record evidence by independent means. Our situation in the moral realm is analogous, perhaps, to the situation we would be in with respect to weather forecasting if we did not have perceptual access to facts about the weather, but instead had to rely on weather-forecasting techniques in order to gauge the accuracy of (other) weather-forecasting techniques.

---

[33] For discussions of calibration in science and elsewhere, see Robert Cummins, "Reflections on Reflective Equilibrium," in Michael DePaul and William Ramsey, eds., *Rethinking Intuition* (Lanham, MD: Rowman and Littlefield, 1998), pp. 113–28; and Allan Franklin, "Calibration," *Perspectives on Science*, v, 2 (Summer 1997): 31–80.

[34] For further discussion of this point, see my "Moral Disagreement and Moral Expertise," in Shafer-Landau, ed., *Oxford Studies in Metaethics*, vol. 3 (New York: Oxford, 2008), pp. 87–108.

Of course, even if I rely exclusively on my own moral judgment in assessing the reliability of others, this does not preclude the possibility that I will conclude that someone else's moral judgment is superior to my own. Suppose, for example, that when you and I disagree, I more often than not come around to your point of view: after further reflection or experience, I typically come to the conclusion that you were correct after all. In this way, I might arrive at the view that your moral judgment is more reliable than my own and resolve to defer to you in cases of future conflict. Still, if this is the kind of route by which one comes to attribute superior moral judgment to someone else, then it is unsurprising that deference in the moral realm seems to be a relatively marginal phenomenon compared to the kind of sweeping deference that we find in many other domains.

In this respect, our attitude toward moral deference seems somewhat similar to our attitude toward deference about philosophy, generally. Here too we have impure cases, in which deference seems relatively straightforward. Consider a case in which you would unhesitatingly defer to a colleague's opinion about some abstruse metaphysical question: you have a generally high regard for her philosophical judgment, she works in the subfield and you do not, and she assures you that the existing arguments in the literature really do favor such-and-such a conclusion. Still, deference about metaphysics, like deference about morality, seems more marginal than deference in many other domains. What could account for this? It is not implausible to suppose that at least part of the explanation is our lack of strong, independent evidence as to which metaphysicians possess reliable judgment. Of course, one might recognize that a particular metaphysician possesses a high degree of philosophical ingenuity and sophistication and, to the extent that one takes such virtues to be positively correlated with reliability about metaphysics, one will take this as some evidence that the individual is reliable. Even on the most optimistic assumptions, however, the possession of such virtues would seem to be relatively weak evidence of reliability compared to the kind of straightforward track-record evidence available in many other domains.[36] Perhaps, then, the explanation for why deference about philosophical truth is relatively marginal has to do with our lack of evidence as to who among the philosophers deserves it.

---

[36] On this point, consider the kinds of disagreement that we find among rival metaphysicians whose philosophical ingenuity and sophistication is not disputed.

The realist might suggest that a parallel explanation holds in the moral domain. She might claim that the best explanation of the datum is not that there are no moral facts, and therefore, no moral experts. Rather, our relative wariness about moral deference is an artifact of the formidable practical obstacles that arise when we attempt to identify or recognize genuine moral experts as such, obstacles that are not present (or present to a significantly lesser degree) in many other domains.

What should we make of this reply? While it seems a clear improvement over the previously considered proposals, there is reason to think that it does not provide a sufficiently deep explanation of the datum. According to the current proposal, there is nothing peculiar about pure moral deference *per se*. Rather, the apparent peculiarity of such deference is an artifact of our epistemic limitations: in the moral realm, we are often simply not in a position to know who deserves deference. The theory predicts that, as our confidence that someone is a genuine moral expert increases, our sense that there is something odd about simply "farming out" our moral beliefs to that person should diminish accordingly. At the limit, we should find the thought of judging that capital punishment is wrong on the basis of someone else's say-so no odder than deferring to him about driving directions. But it is doubtful that our sense that there is something peculiar about pure moral deference would vanish entirely in this way.

To see this, imagine a juror at a criminal trial. The defense calls an impeccably credentialed expert witness who confidently testifies that the accused is insane; the juror thus comes to believe accordingly. But then the prosecution calls its equally well-credentialed expert witness, who with equal confidence testifies to the sanity of the accused. In response, the juror loses his confidence in his earlier belief. He would gladly and without hesitation farm out his opinion concerning the sanity of the accused to one of the experts if only he knew which one deserved his deference in this particular case.[36] According to the current proposal, our sense that there is something peculiar about pure moral deference derives from our implicit recognition that, with respect to morality, human beings generally occupy more or less the same position as the juror. But on reflection, this assimilation seems quite implausible. In particular, our sense that there is something peculiar about moral deference does not seem as contingent as this picture suggests it should be. In contrast, the noncognitivist does have a relatively deep explanation of the datum to offer: on his view, the peculiarity of farming out one's opinions as to the moral facts is due to the nonexistence of such facts.

---

[36] For an interesting discussion of the juror's epistemic plight, see Alvin Goldman, "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research*, LXIII, 1 (July 2001): 85–110.

This objection may not be decisive; in any case, much more could be said. Rather than pursuing this line of thought further, however, I want to turn to a different line of response which the realist might offer, a response which does seem to possess the relevant depth.

V. MORAL DEFERENCE AND AGENCY

Anscombe contends that moral and mathematical teaching are alike in that the primary goal in each case is that the student will learn to *do* something correctly. She argues that, unlike simply believing what one is told to believe, acting correctly requires interpretation:

> There is such a thing as believing what one is told without reflection, consideration or interpretation; doing what one is told is an interpretation and so with doing, however obedient one is, one can hardly escape being one's own pilot... there comes a point where [one] must act and that is the end of listening to advice. But with believing it is otherwise; a man may decline to be his own pilot for certain of his beliefs and altogether rely on authority, without doing anything on his own account to digest and assimilate the beliefs. This would be possible in moral matters only to the extent that his beliefs were idle, without consequences, that is, if they concerned matters that [one] never had to deal with.[37]

Suppose that a child believes the proposition that lying is wrong simply in virtue of deferring to a parent; the child thus lacks any understanding of, or sensitivity to, those features of lying in virtue of which it is wrong. For the reasons given in section III, the realist should admit that, as long as the parent genuinely knows that lying is wrong and the child has no reason to distrust the parent, nothing precludes the child's belief from qualifying as knowledge. Indeed, in the absence of some skeptical view about testimony, it seems that there is no positive epistemic status that the child's belief must lack in virtue of its source. Nevertheless, it is clear that the child's lack of understanding is likely to compromise her ability to apply the relevant piece of knowledge correctly. There are a number of ways in which this is so. First, even if lying is generally wrong, there are, plausibly, exceptional circumstances in which it is not. Assuming that there are some cases in which it is not wrong to lie, one's ability to recognize them as such is presumably not independent of one's understanding why lying is wrong in those cases in which it is. We would thus expect the child to be insensitive to the kinds of considerations that can outweigh or silence reasons against lying in particular contexts. Moreover, if the ability to recognize a possible action as a lie is not wholly independent of sensitivity to those features in virtue of which lies are typically wrong, then this is another respect in which a child whose knowledge is a matter of deference will be ill equipped to apply her knowledge correctly.

---

[37] Anscombe, *op. cit.*, p. 48.

Thus, the realist might follow Anscombe and attempt to accommodate the datum by appealing to the apparent gap between moral knowledge-that and moral knowledge-how. When it comes to morality, the realist might say, understanding why *p* is true (as opposed to merely knowing that *p* is true) puts one in a far better position to apply one's knowledge in particular cases.[38] On this view, an agent whose knowledge depends on pure moral deference resembles the culinary novice who blindly follows the instructions provided by the expert chef. Although the novice might commit the recipe to memory and thus have propositional knowledge of what he or she ought to do at each stage of the process, we should not be surprised if the novice's execution of the task falls short of the expert's. According to this line of thought, our attitude toward moral deference has its roots in our recognition that morality is largely a matter of knowing how, while moral deference typically delivers (at best) knowledge that.[39]

One limitation of this proposal is that it seems to lack the requisite generality: it does not explain the peculiarity of deference with respect to moral beliefs about what to do in particular cases. In the essay from which the above quotation is drawn, Anscombe's concern is with moral teaching, which presumably involves the transmission of moral *dicta* that have a certain degree of generality or abstractness (as opposed to specific injunctions about what to do on particular occasions). Indeed, the line of thought suggested by Anscombe seems most compelling in cases in which there is a substantive gap between what one is told and the concrete circumstances in which one must act: this creates the need for interpretation, and hence, understanding—something which pure moral deference typically does not deliver. But imagine an individual who genuinely comes to believe that *it would be wrong for me to eat this piece of red meat on my plate* in response to having been told as much by one of his dinner companions. Here the need for interpretation seems minimal—it is not as though the individual is in the position of having to apply some general principle. Nevertheless, deference in this case seems no less peculiar on that account.[40] Thus, it does not seem as though the explanation on offer can be the full story.

---

[38] Cf. Nickel, *op. cit.*, p. 261.

[39] The seminal modern discussion of the general distinction between knowledge-how and knowledge-that, as well as the classic arguments for the irreducibility of the former to the latter, are due to Anscombe's Oxford colleague Gilbert Ryle in *The Concept of Mind* (Chicago: University Press, 1949). For recent criticism of the irreducibility claim, see Jason Stanley and Timothy Williamson, "Knowing How," this JOURNAL, XCVIII, 8 (August 2001): 411–44. For more on the distinction in connection with moral testimony, see the brief discussion in Hopkins, *op. cit.*, pp. 618–20.

Still, the idea that the realist might appeal to the intimate link between moral judgment and action seems promising. Consider a somewhat different way of developing it. Again, when one believes that Φ-ing is the right thing to do on the basis of pure moral deference, one typically does not understand why Φ-ing is the right thing to do. (If one did come to see or understand why Φ-ing is the right thing to do, then one's belief would no longer be a matter of pure moral deference.) In section III, we rejected the suggestion that one's failing to grasp the reasons why Φ-ing is the right thing to do would preclude one's knowing that Φ-ing is the right thing to do, or otherwise detract from the status of one's belief. Another possibility, however, is this: if an agent Φ's because of her belief that Φ-ing is the right thing to do, but she does not understand why Φ-ing is the right thing to do, this detracts from the status of her *action*.

Like the proposal just considered, this proposal seeks to accommodate the datum in a way that is congenial to the realist by appeal to the intimate connection between moral judgment and action. But unlike that proposal, the emphasis here is not on the possibility that a lack of understanding will lead one to misapply general moral principles or relatively abstract pieces of moral knowledge when one acts. Rather, the thought is that even when one does what is in fact the right thing to do, acting in the absence of an understanding of why that action was the right thing to do detracts in some (yet to be specified) way from the value of one's action. Of course, the two proposals need not be viewed as competitors, but as supplementing or complementing one another. That is, one might hold that an agent who possesses moral knowledge in the absence of understanding is not only likely to misapply that knowledge in particular cases (and so do the wrong thing), but also such that his morally right actions are less valuable than they otherwise would be.

But what is the respect in which such actions are deficient? Nickel contends that such actions cannot be morally good.[41] His own argument for this conclusion depends on his view, noted above, that in order to fully understand a moral claim, one must possess an independent justification for thinking it is true, or at least have a grasp on its justification conditions.[42] In section III, we noted that the realist will reject any such picture of understanding on pain of compromising his realism. Once this picture has been set aside, is there any reason for thinking that actions of the relevant kind are *morally* deficient?

---

[40] Notice that insofar as we can make sense of a case like this, it is natural to picture the story as involving impure moral deference: the individual takes his dinner companion to have a relevantly similar moral sensibility but to possess some background information about the processes by which the meal was produced, and so on.

[41] Nickel, *op. cit.*

Consider two cases: in each case, the agent genuinely knows that Φ-ing is the right thing to do and responds to this knowledge by Φ-ing. The difference is that in the first case, the knowledge which gives rise to the action is a matter of the agent's having been informed by a reliable moral authority that Φ-ing is the right thing to do in the relevant circumstances, while in the second case the knowledge is based on the agent's own sensitivity to those features of her circumstances which make Φ-ing the right thing to do. In some morally significant respects, the actions seem to be on a par: they will, for example, have the same consequences. But in other morally significant respects they may differ. After all, ordinary moral thought, as well as much moral theory, maintains that it matters a great deal why an agent performs a given action. Perhaps most obviously, there seems to be a morally significant difference between, for example, the shopkeeper who gives his customers the correct change because he is genuinely honest and the shopkeeper who does so solely because it is the most efficient way to maintain an advantageous reputation for honesty.[43] Notice that insofar as pure moral deference does not deliver understanding, pure moral deference does not put one in a position to do the right thing for the right reasons even in those cases in which it delivers genuine moral knowledge. If the child refrains from lying on a given occasion because he knows that lying is wrong but has no grasp on why it is, then his refraining from lying on that occasion is not based on the reasons that there are not to lie.

Thus, we should distinguish between three different agents: (i) the agent who does the right thing for the wrong reasons (for example, the ultimately self-interested shopkeeper), (ii) the agent who does the right thing because it is right, and (iii) the agent who does the right thing for the reasons which make it right. The agent who knows that Φ-ing is the right thing to do on the basis of pure moral deference is in a position to do the right thing because it is right, but he is not in a position to do the right thing for the underlying reasons which make it right. To the extent that we attach a greater value to the latter, we have reason to be dissatisfied with pure moral deference, even in those cases in which it delivers genuine moral knowledge.[44]

---

[42] On the exegetical issue, see note 30 above.

[43] The example, of course, is from Kant's *Grounding for the Metaphysics of Morals* (Indianapolis, IN: Hackett, 1981 [1785]).

In attempting to accommodate the datum in this way, the realist appeals to a broadly Aristotelian conception of right or virtuous action. In *Nichomachean Ethics* II.4, Aristotle poses a puzzle for his own account of moral development.[45] According to that account, one must perform just actions in order to become a just person. However, it seems as though someone who performs just actions is *already* a just person, in the way that someone who successfully produces the kind of thing at which a given craft aims is already a craftsman of the relevant kind. In response to this puzzle, Aristotle offers two replies, each of which is suggestive in the present context. First, he distinguishes between conformity and understanding: even with respect to the crafts, one might produce the very kind of thing at which the craftsman aims without yet being a craftsman oneself. This would be the case, for example, if one was simply following the instructions of someone else, as opposed to acting on one's own understanding. Moreover, Aristotle holds that there is a crucial disanalogy between the products of a craft and virtuous action. With respect to the products of a craft, it is enough that they are in the right state once they have been produced (even if they came to be in that state by chance, or at the hand of someone who was blindly following the instructions of another person). In assessing the quality of what has been produced, no further question needs to be asked about the state of the person who produced it. However, things are otherwise with respect to just actions: in order for a "just action to be done justly" it is crucial that the agent himself acts in a certain state, a state that (among other things) requires understanding.

---

[44] Cf. Hills on the conditions for "morally worthy action," *op. cit.*, pp 113–17

Might the attractiveness of this realist response depend in part on subtle issues about what provides us with reasons for action? Consider, for example, Thomas Scanlon's "buck-passing" account of goodness, as defended in his *What We Owe to Each Other* (Cambridge: Harvard, 1998). According to this account, to describe an action as good is not to attribute some additional, reason-providing property to it. Rather, to describe an action as good is to attribute to it the higher-order property of possessing some lower-order property or properties that provide reasons to perform it. Thus, if a given action is good because it is a kind action, it is not, strictly speaking, the fact that the action is good which counts as a reason to perform it; rather, it is the fact that it is kind which gives one such a reason Scanlon contrasts the buck-passing account with a view about reasons for action that he attributes to G. E. Moore. According to that view, it is not the fact that the action is kind that gives one a reason to perform it, but rather the fact that the action has the property of being good So perhaps the current proposal would hold more appeal for those realists who accept a theory of reasons such as Scanlon's, as opposed to the Moorean alternative: on the Moorean view, but not on Scanlon's view, an agent who knows only that a given action would be morally good is in a position to act on those reasons which speak in favor of the action

[45] In this paragraph, I have relied on Terence Irwin's translation of the *Nichomachean Ethics* (Indianapolis, IN: Hackett, 1985).

Notice that a realist who pursues this strategy is not committed to the view that an agent who does the right thing because it is right fails to be morally praiseworthy. While an agent who does the right thing for the wrong reason would seem to be an inappropriate object of praise, there are many cases in which an agent who does the right thing because it is right would seem to deserve praise, even if he acts in ignorance of the reasons which make that action the right thing to do. (Consider especially cases in which the agent incurs considerable and foreseeable costs by performing the action.) It is not, after all, as though such an agent displays a lack of concern for morality. On the contrary, it is his very concern for morality that motivates him to act as he does.[46] The point is rather that, even granting that such an agent might very well deserve moral praise, he or she still falls short of an important ideal associated with moral agency: that of doing the right thing for the reasons that make it right. Thus, the realist might seek to accommodate the datum by claiming that our relative aversion to moral deference is due to our recognizing that someone who relies on moral deference for her moral views will inevitably fall short of this ideal. By contrast, when a person defers about some empirical matter, her doing so does not similarly frustrate the achievement of any ideal associated with agency. Thus, deference about the weather, geography, or some esoteric scientific matter strikes us as unproblematic in a way that moral deference does not.

What should we make of this proposal? I believe that it is the most promising of the strategies that we have considered. Nevertheless, it too faces challenges. While a full development and defense of it is beyond the scope of this paper, I will close by presenting two of the more pressing challenges and gesturing at how a realist might respond.[47]

The first objection is that, in providing an agency-based explanation of the datum, the proposal overestimates the extent to which our moral judgments are tied to action. As is often noted, we seem to make moral judgments that lack any practical import. For example, we might judge that some ancient practice was morally barbaric, even though that practice is not a live option for us, and even though the fact that it is not a live option for us does not depend on moral considerations (that is, even if we changed our minds about its moral status, there is no possibility that we would engage in it or any similar practice). On the face of it, the current proposal would seem to predict that pure moral deference about such matters should strike us as unproblematic, or at the very least, significantly less problematic than deference about matters of potentially practical significance. But it simply does not seem as though this is so. Offhand, my deferring to you about the moral status of abortion seems no less and no more peculiar than my deferring to you about the moral status of some long-abandoned ancient ritual.

---

[46] Perhaps there is a respect in which the ignorant agent who does the right thing because it is right is less open to criticism (and potentially, more deserving of praise) than the agent who does the right thing because it is right even though he knows why it is right. The latter agent, unlike the former, is in a position to do the right thing for the reasons that make it right but fails to be moved by those reasons, falling back instead on the fact that the action is right. The latter agent thus seems a more natural target for the charge that he is "fetishizing" morality in an objectionable way. On "moral fetishism," see Smith, *op. cit.*

[47] For further development and defense, see my *Knowledge in the Moral Domain* (manuscript).

A second objection proceeds from the apparent possibility of an agent's deferring to another person not only about the claim that a particular course of action is right, but also about the underlying reasons that make it right. For example, it seems that in principle I could learn from you not only that a given action is the right thing to do, but also that that action is the right thing to do because it is a kind action. It appears, then, that I could go on to perform the action for the reasons that make it right even though I have deferred to you about what those reasons are. Insofar as there seems to be something peculiar about my deferring to you in this way, this apparently cannot be explained by the claim that such deference frustrates the ideal of doing the right thing for the reasons that make it right.

How might the realist sympathetic to the proposal respond to these objections? A natural strategy for overcoming the first objection might involve attempting to show that our attitude toward moral deference in cases in which nothing practical is at stake derives from our attitude toward such deference in other, allegedly more fundamental cases. In response to the second objection, the realist might argue that an agent can know that a given action is the right thing to do, and even know that it is the right thing to do because it has such-and-such features, yet still not be in a position to perform the action on the basis of the relevant reasons. More precisely, the realist might argue that in order to act on the basis of the reasons which make the action the right thing to do, the agent must appreciate those reasons as reasons and not merely possess propositional knowledge that they require the action. Compare: one might know both that some mathematical claim is true and that some collection of propositions makes up a canonical proof of it, but not yet see the collection of propositions as a proof of the claim. This might be the situation of a student of mathematics whose knowledge is a matter of pure deference to her teacher. In that case, the student is not yet in a position to believe the claim on the basis of the proof in the way that the teacher is. Similarly, the realist might say, an agent who is dependent on pure moral deference for her knowledge that an action has certain right-making features is typically not in a position to perform the action on the basis of those reasons, in the way that someone who genuinely grasps the connection between the two is.

It remains to be seen whether the current proposal is ultimately defensible against these and other objections; perhaps it could provide a satisfying response to the puzzle if spelled out in full detail. Alternatively, perhaps some other realist strategy might be vindicated. In the absence of some such vindication, however, the fundamental puzzle for the realist must be regarded as unsolved.

SARAH MCGRATH

Princeton University

CENTRAL LIBRARY UNIVERSITY OF CALCUTTA CORE CAMPUS

HOW TO BE A PESSIMIST ABOUT AESTHETIC TESTIMONY[*]

Testimony is an important source of knowledge. In many areas, a good deal of what an individual knows, she knows on the word of others. This holds even if we concentrate (as we will) on pure cases of testimony—those in which the audience *H* learns that *p* on the basis that her informant *T* claims that *p*, and independently of any evidence that *T* offers for that claim. Indeed, in general, where there is something to know, testimony provides a legitimate way to acquire that knowledge. Of course, one cannot believe just anybody. Gullibility is to be avoided, and informants who themselves lack knowledge cannot, perhaps, instill knowledge in their audience—even if what they say is true. But, provided one is suitably careful, on most issues taking the word of others is a legitimate way to come by belief.

There are, however, putative exceptions. Aesthetic matters provide a prominent example. If you tell me that some film you have seen is excellent, it is far from clear that I can legitimately adopt your view. Your testimony might motivate me to watch the film, but it does not give me the right to the belief that the film is excellent. I will not have that right until I have seen the film myself. Or so, at least, some think.[1] Others disagree.[2] They suggest that there need be nothing wrong with taking testimony on aesthetic matters. As before, one must act responsibly in choosing whom to believe. Perhaps aesthetic matters require

---

[*] I thank audiences at the annual meetings of the American and European Societies for Aesthetics, at the Weggis conference on Normativity, at the Auburn conference on Beauty, at University College London, and at the University of Leeds. I am particularly grateful to Fabian Dorsch, John Gibbons, Keren Goredeisky, Mike Martin, Aaron Meskin, David Owens, and Gianfranco Soldati.

[1] See Roger Scruton, *Art and Imagination A Study in the Philosophy of Mind* (London: Methuen, 1976), chapter 4; Richard Wollheim, *Art and Its Objects*, 2nd ed. (New York: Cambridge, 1980); Philip Pettit, "The Possibility of Aesthetic Realism," in Eva Schaper, ed., *Pleasure, Preference and Value Studies in Philosophical Aesthetics* (New York: Cambridge, 1983), pp. 17–38; Michael Tanner, "Ethics and Aesthetics Are — ?" in José Bermúdez and Sebastian Gardner, eds., *Art and Morality* (New York: Routledge, 2003), pp. 19–36; Robert Hopkins, "Beauty and Testimony," in Anthony O'Hear, ed., *Philosophy, the Good, the True and the Beautiful* (New York: Cambridge, 2000), pp. 209–36, and Hopkins, "Kant, Quasi-Realism, and the Autonomy of Aesthetic Judgement," *European Journal of Philosophy*, ix, 2 (2001): 166–89.

[2] See Malcolm Budd, "The Acquaintance Principle," *British Journal of Aesthetics*, xliii, 4 (October 2003): 386–92; and Aaron Meskin, "Aesthetic Testimony: What Can We Learn from Others about Beauty and Art?" *Philosophy and Phenomenological Research*, lxix, 1 (July 2004): 65–91.

0022-362X/11/0803/138–157

greater caution than some others. But there is nothing wrong in principle with acquiring aesthetic knowledge in this way.

Call *pessimists* those who deny the legitimacy of using testimony as a source of aesthetic belief, and *optimists* those who accept it. If their debate were merely over how often, or under what conditions, aesthetic matters admit of testimony, it would be hard to marshal. A range of positions would be available, forming a sliding scale between thinking that testimony is never a legitimate source of belief and thinking that it is always so, or at least as often as in other matters. Fortunately, the issue can be focused by appeal to the idea of a difference in kind. Can I legitimately take another's word on an aesthetic issue? Pessimists say "no": aesthetic matters differ in kind, in this respect, from most others. Optimists disagree: one can take aesthetic testimony, and any difference here is merely one of degree.

Aesthetics is not the only area in which something like this debate occurs. There are pessimists about moral testimony, and pessimists about testimony in mathematics. It is easy to imagine pessimists about comedy. Although my topic is pessimism about aesthetic testimony, and that is what I will mean by "pessimism," it will sometimes prove illuminating to compare it with analogous positions elsewhere.

Whatever the area, the disagreement between pessimists and optimists turns, in key part, on conflicting intuitions over whether in particular cases it would be legitimate to take the testimony one is offered. Such conflicts are hard to resolve. Either side struggles to persuade the other that its description of the case gives the wrong result, and stalemate threatens. But even if we are stuck with this disagreement, neither side should simply ignore the intuitions it rejects. Any account of aesthetic testimony should seek to explain why intuitions divide in this way. One of my goals is to offer such an explanation. My main ambition, however, is to explore the forms that pessimism might take, and in particular the form that in some ways is most promising.

I

Although there has been quite extensive discussion of pessimism, it has gone unnoticed that it might take two forms. According to the pessimist, something prevents testimony from being a legitimate source of aesthetic knowledge. The question naturally arises: what? There are two possibilities. As a rough first pass, we might say that the obstacle to reliance on testimony in aesthetic matters might be epistemic or it might be aesthetic. On the first sort of view, aesthetic testimony fails to meet the conditions that, quite generally, govern testimony's ability to act as a source of knowledge. Something or other prevents our aesthetic claims from offering knowledge to those to whom they are directed. Since the epistemology is wrong, there is no knowledge to be had from such testimonial transactions, and that is why it is illegitimate for the listener to let what she is told shape her aesthetic belief. The alternative account, in contrast, accepts that, at least in favorable circumstances, aesthetic claims can meet every condition governing whether testimony yields knowledge. The obstacle to taking aesthetic testimony instead lies in some further norm specific to the aesthetic case. Since the epistemology is (sometimes) right, testimony does offer us aesthetic knowledge. Some other norm, however, renders it illegitimate to make use of that resource.[3]

So far, we know almost nothing about what this further norm might be. All we know is that it is not one of the norms governing whether testimony in general is a source of knowledge. It is possible, therefore, that the content of this norm is in some way epistemic—for instance, that it limits legitimate aesthetic beliefs in terms of the methods by which they are formed, or the nature of our justification for them. And it is possible that the sources of this norm, the considerations that explain why it holds, do not in any clear sense count as aesthetic. Below we will see that both possibilities prove real. It is in some ways misleading, then, to think of the contrast as between epistemic and aesthetic obstacles to aesthetic testimony. Better, I suggest, to characterize it as between *Unavailability* and *Unusability*, between accounts on which testimony does not make aesthetic knowledge available, and those on which, though it does, it is not acceptable to make use of that knowledge to form belief.

II

If pessimism can take either form, which is more plausible? *Prima facie*, there are two grounds for preferring Unusability.

The first is that none of the obvious Unavailability accounts succeed. Such accounts seek to specify how exactly aesthetic testimony fails to make knowledge available. There is no space here to discuss these views in the detail they deserve. Still, since some of these accounts will occur to many a reader, it may help to have the briefest sketch of how they fail. One radical position holds that testimony cannot be a source of aesthetic knowledge for the simple reason that there is no aesthetic knowledge. Aesthetic judgments cannot be true, and perhaps are not even truth-apt. Instead, we should adopt some sort of error theory or expressivism concerning them.[4] The problem this explanation faces is that it is in tension with other things the pessimist will say. Pessimists deny us the right to form aesthetic beliefs on testimony. What do they offer us instead? They exhort us to settle matters for ourselves. In the aesthetic case, this usually involves exposing ourselves to whatever object the judgment concerns, and trying to come to our own view of its qualities and quality. But what is the point of settling the issue for ourselves, if there is nothing for aesthetic testimony to pass on? If our own investigations are valueless, there is no point conducting them; and if they bring some benefit, why can that not be passed on via testimony?

---

[3] Hence only some pessimists about aesthetic testimony are skeptics, if that means denying that it yields knowledge.

[4] For error theory, see J. L. Mackie, *Ethics: Inventing Right and Wrong* (New York: Penguin, 1977). For expressivism, see Scruton, *op. cit.*; Simon Blackburn, *Spreading*

Another Unavailability account appeals to the extent of disagreement in aesthetic matters. Where informants disagree, some must be unreliable. The more disagreement there is in an area, the more unreliable we must take informants on those matters to be. Aesthetic testimony fails to deliver knowledge because it is too difficult to identify reliable informants.[5] The problem for this account is to explain the failure of aesthetic testimony without mandating agnosticism. The more widespread is disagreement, the less reason there is to think that *anyone* is a good judge in the matter in hand—oneself included. Thus it is unclear how this account justifies resisting testimony without requiring one to eschew aesthetic belief altogether.[6]

The second reason to prefer Unusability accounts is that under the right conditions testimony simply must make aesthetic knowledge available (assuming there is any such knowledge). One way in which testimony can be a source of knowledge is if the recipient is able to form a Humean argument for believing what she is told. Such an argument reasons to the truth of the informant's claim that p from the type of informant and the type of topic on which he is offering a view; the past correlation, as experienced by the recipient, between testimony by such informants on such matters and the truth of their claims; and the prior probability, for the recipient, that p. A Humean argument is thus an instance of inductive reasoning. Assuming that inductive inference can yield knowledge, that is, setting aside skepticism about induction, there are conditions under which a Humean argument, as applied to aesthetic testimony, will be knowledge-yielding.

---

the Word: Groundings in the Philosophy of Language (New York. Oxford, 1984), and *Essays in Quasi-Realism* (New York: Oxford, 1993).

[5] Meskin, *op. cit.*

[6] Thus the two Unavailability accounts suffer analogous defects. The difference is that, while the first threatens to deny that there is anything valuable for one's deliberations to yield, the second threatens to deny that those deliberations have any chance of yielding that good.

We merely need the claim testified to to be true, the experienced correlation to come out sufficiently high, and the subject's prior probability that p not to be too low. Even if such conditions never actually obtain, surely they could. What will the pessimist say about that possible situation? Of course, anyone with doubts about the legitimacy in general of taking testimony on aesthetic matters could at this point fold. There is nothing wrong with taking testimony under conditions in which it, coupled to the availability of a sufficiently strong Humean argument, would yield knowledge. But I suspect that many of a pessimist persuasion will feel that there is something wrong with taking aesthetic testimony, even under conditions as favorable as this. For such folk the only resort is Unusability. *Ex hypothesi* knowledge is made available by this testimony, yet taking it remains illicit. The source of that illegitimacy can only lie in some norm other than those determining whether knowledge is available. Since there is nothing special about this case other than the degree to which it favors knowledge, that other norm presumably stands in the way of all aesthetic testimony.

III

Those drawn to Unavailability will no doubt object to both these reasons for preferring Unusability. To the first, they will respond that there is no clear advantage here until some satisfying Unusability account has been developed. Perhaps it is difficult to provide a satisfying detailed view of either form. If so, while pessimism might be in trouble, neither variant fares better than its rival. I accept this reply. There is as yet no proven benefit to Unusability. Later I will explore some of the ways in which it is indeed unclear what the details of a decent Unusability account might be.

The response to the second point in favor of Unusability is more complex. The possibility of an ideal Humean argument shows that aesthetic testimony can, in principle, yield knowledge. But it does not show that such testimony, even under ideal conditions, exhibits the epistemology of testimony in general. For some think that testimony does not, in general, yield knowledge by Humean means. In the standard case, H does not learn from T that p in virtue of the fact that a persuasive Humean argument for p is available to H. Rather, H has some default right to believe what her informants tell her. That default can be overridden, if the conditions are unfavorable—for instance, if the informant gives grounds for suspicion. H must conduct herself responsibly, if she is to retain that right—for example, she must be sensitive to such indications that T is untrustworthy. And, even where H has the right to believe T, her acquiring knowledge turns on things beyond her control. In particular, she will acquire knowledge only if *T* has it to pass on—only, that is, if *T* knows the claim to which he testifies. But if all this is in place, *H* inherits the knowledge that *T* offers her. Given the right framing conditions, testimony offers a way to transmit knowledge from one person to another, regardless of the availability to the recipient of an argument for believing what she is told. The case of the perfect Humean argument shows nothing about whether this form of epistemological interaction is possible in aesthetics. It shows that aesthetic testimony can, under ideal conditions, yield knowledge, but not that it can do so via the means standard for testimony on other matters. The case thus leaves it open that aesthetic testimony differs in kind from testimony on other matters, and that the difference lies in its epistemology. Unavailability has not been disproved.

The idea that testimony does not, in general, offer knowledge via Humean means is attractive. Alternative accounts invoking the idea that, given the right framing conditions, knowledge is transmitted to the recipient, have been moderately well worked out.[7] Further, the distinction between Humean and Transmission accounts of testimony promises another benefit to the proponent of Unavailability. It offers her a way to explain the conflict in intuitions over aesthetic testimony with which we began. She can interpret the intutions that drive optimism as reflecting the fact that aesthetic testimony can, in principle, yield knowledge, via Humean inductive arguments. And she can do this while taking the pessimistic intuitions as reflecting an awareness that aesthetic testimony nonetheless fails to reproduce the epistemic situation that obtains for testimony on other matters. Even when it does make knowledge available, it does not do so by the means standard for other sorts of testimony.

Nonetheless, I am not persuaded by this response. First, it is not clear that appealing to the distinction between Humean and Transmission models preserves more than the letter of Unavailability pessimism. For, while the resulting view is able to claim an epistemic difference in kind between testimony on aesthetic and on other matters, it no longer tells us what is wrong with the former. Claiming that such testimony does not make knowledge available by transmission leaves open that it does so by Humean means. After all, nothing the view has said provides reason to deny that appropriate Humean arguments are widely available and that such arguments are legitimate sources of aesthetic belief. If so, there is nothing wrong with taking our aesthetic beliefs from others. Unless supplemented by quite distinct claims concerning the difficulty in aesthetic matters of providing Humean arguments, the account thus does not explain what is wrong with aesthetic testimony. Of course, the view might nonetheless provide the correct description of the epistemic facts. Perhaps aesthetic testimony does make knowledge available, but only by Humean means. However, a second point is that we have yet to be given any reason for believing this. Why think that aesthetic testimony fails to operate in the way Transmission accounts describe? Until we have an answer, the response has been sketched, not developed.[8] It is true that the view already promises to do some work, in explaining the conflict in intuitions over aesthetic testimony. Perhaps that gives us some reason to believe it. But my final point is that this explanation is less satisfying than one in terms of Unusability.

---

[7] See, for instance, Tyler Burge, "Content Preservation," *Philosophical Review*, CII, 4 (October 1993): 457–88; John McDowell, "Knowledge by Hearsay," in *Meaning, Knowledge, and Reality* (Cambridge: Harvard, 1998), pp. 414–43

Is it really plausible that our conflicting intuitions reflect a dim awareness of differing epistemologies of testimony? At least one of the epistemologies, that in terms of transmission, is somewhat recherché. We might expect our intutions to track whether testimony yields knowledge; it is another thing to claim they track the various means by which that knowledge is yielded. Contrast what Unusability views will say about the conflict of intuitions. Such views need not be at all sniffy about the prospects for aesthetic testimony's making knowledge available. They can be open-minded about the range of conditions under which such testimony might do this, and about how often those conditions actually obtain. So they can acknowledge that something important is right in the intuitions that drive optimism. What optimists are picking up on is that there is nothing epistemologically wrong with aesthetic testimony. What the intuitions driving pessimism reflect is that there is something wrong with it, nonetheless. Some further norm renders it illegitimate to make use of the knowledge such testimony offers us. This elegant account of the conflict demands, of those in the grip of the intuitions, rather less in the way of philosophical insight than its rival. To my mind, that makes it more plausible.

Unusability thus has two advantages over Unavailability. It offers a more straightforward way to accommodate the thought that under some conditions aesthetic testimony must make knowledge available. And it offers a more plausible explanation of why intutions over aesthetic testimony conflict. In the rest of this paper, I propose to examine the nature of, and prospects for, Unusability accounts.

---

[8] In "Beauty and Testimony," I offered the missing argument for the response. However, at the time I was not alert to the distinction between Unavailability and Unusability. The argument I offered fails to support the response over the quite different account of the phenomena in terms of Unusability.

IV

I begin with a threat to the very coherence of such views. One might wonder whether it is possible for belief to be governed by norms of the kind Unusability postulates. Belief aims at truth, and aspires to the status of knowledge. Epistemic norms determine whether it hits this target. How can belief be governed by nonepistemic norms, norms that govern something other than whether it counts as knowledge? If someone's testimony offers one knowledge, how can one fail to mold belief to fit? *Ex hypothesi* the claim being made is true and the situation is such as to put one in a position to know it. What room is there for a norm governing belief that requires one to refuse the knowledge thus offered? It seems that knowledge is not merely a goal of belief, but *the* goal. Epistemic norms are thus the only ones that should govern its formation.

These thoughts motivate a rather different position on testimony from any considered above. Belief may be governed only by epistemic norms, but assertion is certainly subject to other demands. In deciding who is responsible for the recent mishap, one should seek to be responsive only to factors determining whether one's conclusion will count as knowledge. One's assertions on the matter, in contrast, may be ruled by quite other considerations, including tact, prudence, and concern for the feelings of others. It is what we go on to say, not what beliefs we form, that is governed by norms other than those determining whether our belief would count as knowledge. Thus, if the epistemology is right, we can legitimately let testimony guide aesthetic belief; it is assertions of that belief that are proscribed by some nonepistemic norm.

This position is too weak to count as pessimism. Nonetheless, it fits some of the pessimistic intuitions perfectly well. For some of the most striking phenomena suggesting that there is something problematic about aesthetic testimony involve conversational implicature. Suppose I do take your word on the excellence of some film. There is something strange in my then saying to a third party "*Doubt* is great," yet the oddness is dispelled if I preface that claim by "According to a friend of mine, ..." One explanation of this is that the former claim conversationally implicates, as the latter does not, that I have come to my own view on the merits of the film. Since implicature essentially concerns assertion, not belief, the new account is perfectly positioned to capture these phenomena.

However, I think the pessimist can draw on intuitions that concern belief.[9] If so, the moderate position just described does not do justice to the full range of pessimist intutions. More importantly, those attracted to Unusability are not forced to retreat to that position. For the thoughts that motivated that retreat are misplaced. Belief can be governed by nonepistemic norms. Or so I shall now argue.

---

[9] See Hopkins, "Beauty and Testimony."

First, notice that the Unusability pessimist is not claiming that one can form belief in response to nonepistemic norms. That would be to form it in part without a view to its truth or its status as knowledge, and that may indeed be inconsistent with the goals constitutive of belief. Rather, the claim is that responsiveness to nonepistemic norms can lead one to *resist* forming belief. By governing belief only in this direction these norms leave it as a necessary condition on one's beliefs that one takes the propositions in question to be true and takes oneself to stand to them in epistemically satisfactory relations. All they block is the idea that meeting these conditions is sufficient for a proposition to merit one's belief.

Those skeptical about the coherence of norms of Use may consider this reply too limited. True, it helps that the suggestion is not that belief should be formed with a view to something other than how things are. That deals with the most serious potential objection to the view. But doesn't a significant obstacle remain? Isn't belief's role *nothing more* than to reflect the facts? If so, it can be governed *only* by epistemic norms. So one may remain tempted to think there is no room for norms of the kind Unusability views postulate.

It is as well, then, that a second response to the challenge is available. At least some of our practices of belief formation involve norms of precisely the kind Unusability postulates. For examples to be persuasive, we need to get away from the particular subject matters—aesthetics, ethics, mathematics, and perhaps others—that give rise to the debates between pessimists and optimists, and between pessimists of different stripes. But such cases can be found. For we find norms of the kind we seek in many areas. All that is required is that there be room for one particular notion of expertise.

Although the notion of an expert is multi-faceted, one element that sometimes comes to prominence is the idea that an expert is someone who *ought to settle for herself questions in her domain of expertise*. As philosophers, we should be familiar with this idea, for it governs our own professional lives. It is part of our job to think philosophical issues through, and to reach our own judgments on them. It is consistent with pursuing this goal that we consider others to be as good philosophers as we are, or better. If we do think this, we might think that their testimony on a philosophical matter offers us knowledge. But we might still resist forming belief on that basis, and continue working through their reasoning for ourselves. Only if we come to accept their arguments, and to see the truth of their claims, will we adopt the views in question. Perhaps the idea of philosophical knowledge makes you uneasy. If so, just change the example. Think of chemists, or logicians, or historians, or whatever you like. Of course, for any such specialist there will be some topics, relevant to her field of expertise, on which she does think it acceptable to take testimony. The point is that for many topics, those forming the core of that field, she will not think it so. And this, even though she may consider the judgment of her peers to be worth as much as hers, or more. In such circumstances, she resists testimony, even though it offers her knowledge, and she takes it to do so.

I think the case of experts provides a clear example of belief formation governed by a nonepistemic norm. Since these folk resist taking testimony even when they judge that it offers them knowledge, they cannot be responsive only to norms that determine whether their prospective belief counts as knowledge. Since there is nothing incoherent in this practice, which is familiar to us all, there can be nothing incoherent in the idea of a nonepistemic norm that governs belief, a norm of Use.

Two things might be said in reply. The first is that the norm here itself has an epistemic justification. The reason experts are required to settle for themselves matters in their field is that that way we minimize the risk of error across the interested community as a whole. By holding out until we fully appreciate the weight of the arguments supposedly establishing a position, each of us forces the advocates of those arguments to re-examine and refine them, raising the chance that any hidden errors are brought to light. All this is true. It is also true that, in the aesthetic case, if there is a nonepistemic norm governing aesthetic belief, it may not find a similar epistemic justification. But that is irrelevant to the point made by the case of experts. Whatever the justification for the norm governing expertise, it remains nonepistemic in the sense at stake. It is not a norm that determines whether the belief the recipient is offered would count as knowledge. It might increase the chances, across the community as a whole and the beliefs they collectively hold, that more of those count as knowledge. It thus bears indirectly on the status of any belief testified to. But the case goes through even when the status of the belief currently being offered is not up for grabs. The recipient need have no real doubt that her esteemed colleague's view amounts to knowledge, and yet feel she ought not to take her word on the matter.

The second reply offers an alternative way to describe the case. Being an expert does not oblige one to resist knowledge that testimony offers. Rather, it raises the bar for what, in the case of one's own beliefs, counts as knowledge. The laity might gain knowledge from the colleague's testimony, but our expert cannot. Thus the norm to which the expert responds is epistemic after all. It is just that what counts as knowledge for some (the lay recipients of the testimony) does not count as knowledge for others (the original expert herself).

This proposal amounts to a form of contextualism about knowledge. What is distinctive here is that the context in question is wholly constituted by the professional role a person occupies. We need not worry whether in general contextualist views are plausible. Even if so, the redescription here either misrepresents the facts, or concedes the key point that Usability seeks. Contextualism is usually taken to be the idea that the degree of justification or warrant required for knowledge varies with context. If we stick to that interpretation of the position, it misdescribes the case in hand. For, in resisting testimony, our expert need not believe that thinking the matter through for herself will result in her having greater justification or warrant for the view upon which she settles. Her modesty might run to the thought that, not only does her colleague's testimony offer her knowledge, but it offers her as great a warrant or justification for the belief testified to as she can reasonably hope to attain by working things out for herself. Even if she thinks thus, she can coherently resist the testimony, on the grounds that she ought to settle the matter independently. Suppose the contextualist accepts this, and admits that context does not merely raise or lower the bar for knowledge, in terms of the degree of support a belief requires to count as such; it also dictates directly what methods of belief formation are appropriate. Then the contextualist description of the facts is a mere terminological variant on that offered by the Unusability pessimist. The norm the contextualist proposes, like that mooted for aesthetic testimony, dictates that one can legitimately form belief only by settling the matter for oneself. It has the same consequences for our practices of forming belief as that postulated by the Unusability view. But then there is after all nothing incoherent in the idea that belief formation might be governed in this way. It matters not one jot whether we follow this contextualist in labeling the norm with these consequences epistemic. If his description of the case of experts is coherent, so is the Unusability account of aesthetic testimony.[10]

---

[10] So the dialectic has played out as follows. Confronted with a threat to the coherence of the idea of a norm of Use, we appealed to a case, that of experts, which seems both to involve such a norm and to be coherent. The contextualist tries to redescribe that norm as one of Availability. But that description, if it is to be accurate, involves that parallel context being governed by a norm, however labeled, of the precise nature

V

Prima facie then, Unusability views are coherent, and have at least some attractions over the rival form of pessimism. However, they face several questions. The first is what the nonepistemic norm governing aesthetic testimony might be. Its upshot is clear: even when such testimony offers one knowledge, one ought not to exploit that resource. But what is the norm that has this as consequence?

I consider two candidates. To reach the first, consider the analogous position in ethics. Anyone pessimistic about the legitimacy of taking our moral beliefs from others will find that the distinction between Unavailability and Unusability can be drawn there too, and that there too Unusability has certain advantages over Unavailability.[11] Confronted with the parallel question to that now before us, such a pessimist might offer the following as the nonepistemic norm governing moral testimony:

> *The Requirement (for morals)*: Having the right to a moral belief requires one to grasp the ***moral*** grounds for it.

This has a natural analogue in mathematics. Note that the norm is epistemic in content. As I warned earlier (section I), that may be a feature of some norms of Use. The point is that it is not epistemic in the sense at issue: it is not a norm determining whether any belief adopted on the basis of the testimony before one would count as knowledge. This is not explicit in the norm itself; but that is the intent behind it.[12]

Thus we might hope to draw out the links between the situation concerning aesthetic testimony and that holding in these other areas by offering the following as the norm governing the former:

> *The Requirement (for aesthetic matters)*: Having the right to an aesthetic belief requires one to grasp the ***aesthetic*** grounds for it.

Unusability postulated. The challenge to coherence has thus been met. There is a further question. If we can describe the norm either as one of Availability or one of Use, then have we characterized the distinction between those two approaches adequately? There may be more work to do here. However, my main goal is to explore the view of the facts offered by Unusability. If that view can, given certain assumptions, be expressed as Unavailability, that does not prevent me from meeting that goal.

[11] I explore this issue in Hopkins, "What Is Wrong with Moral Testimony?" *Philosophy and Phenomenological Research*, LXXIV, 3 (May 2007): 1–24

[12] It would perhaps be undesirable for norms of Availability and Use to find identical expression. The best way to avoid this is not to change the Requirement, but to make the difference explicit in the corresponding norm of Availability. It might read, "If a moral belief is to count as knowledge, one must grasp the moral grounds for it."

For brevity, I will call this the "Requirement." How does it exclude forming aesthetic belief on testimony? We saw that there are two main models for how testimony operates. Whichever operates in a given case, testimony cannot meet the Requirement. On the Humean model, although the subject has grounds for her belief—provided by the Humean argument—those grounds are not aesthetic. Far from concerning features salient to the aesthetic aspects of the object judged, they concern the type of informant, and the correlation between such informants and the independently established truth of their claims on aesthetic matters. And on the Transmission model, the subject has no grounds at all for the belief she adopts. Rather than her right to the belief depending on the availability to her of some argument, or some other set of reasons, she is by default entitled to adopt the view she is offered, provided certain framing conditions are met. On either model, then, the Requirement promises to proscribe aesthetic testimony.

However, for this to go through, we need to know what counts as an aesthetic ground for an aesthetic claim. I will simply assume that we have a sufficient grasp on the notion of the aesthetic to know (by and large) which beliefs are aesthetic and which not. But what is an aesthetic ground for such a belief? And what counts as grasping such a ground? These questions are not easy to answer. A plausible first stab would be this: to grasp the aesthetic grounds for an aesthetic belief is to grasp those facts in virtue of which the belief is true. If a pot, for instance, is beautiful, then (according to the Requirement) having the right to believe it so is to know what makes that pot beautiful: its shape, its coloration, the texture of its surface, and so forth. Presumably the demand is not just that the subject knows the pot to have those features, but that further she has some appreciation of how those facts bear on the pot's being beautiful. Perhaps these suggestions raise as many questions as they answer. Even so, I will assume that we have sufficient understanding of the Requirement to proceed.

VI

The other candidate for a norm of Use is familiar from the literature on aesthetics. It is the norm Richard Wollheim dubbed the "Principle of Acquaintance":

> *Acquaintance Principle*: Having the right to an aesthetic belief requires one to have experienced for oneself the object it concerns.[13]

---

[13] See Wollheim, *op. cit.*, p. 234. My formulation of the Principle differs from Wollheim's. For present purposes, the differences are unimportant.

In the case of the belief that the pot is beautiful, the Acquaintance Principle demands that one have experienced—presumably seen, touched, or both—the pot. Only then does one have the right to that belief. Since the testimony of another cannot offer one the chance to experience the pot, the Principle excludes taking testimony as a legitimate source of aesthetic belief.

Wollheim is not the only person to have discussed this Principle.[14] Every discussion, sympathetic or otherwise, has assumed the Principle to be an epistemic norm, one determining when an aesthetic belief counts as knowledge. We now see another possibility. The Principle might be taken as a norm of Use. Even when some method of forming aesthetic belief offers one knowledge, the Principle denies one the right to the resultant belief unless one has experienced the object for oneself. It is this interpretation of the Principle that interests me.[15]

How do the Acquaintance Principle and the Requirement relate? There are powerful pressures working to draw the two together in terms of which instances of aesthetic belief they license. However, the two are distinct. They differ in content: the Requirement demanding that we grasp those facts in virtue of which our aesthetic beliefs are true; the Principle, that we experience for ourselves the objects of those beliefs. And, even if they will often be met together, the overlap in which beliefs they license will never be complete. For, while both exclude forming aesthetic belief on testimony, only the Principle has implications for another important potential source of aesthetic belief.

The idea of a principle of taste was first clearly formulated by Kant.[16] Such a principle is a universal generalization, to the effect that anything with some feature *F* has some aesthetic feature *G*. (In more sophisticated variations on the idea, the principle might merely claim that whatever is *F* is *pro tanto*, or tends to be, *G*. This and other subtleties need not detain us.) The interest of such principles lies in the possibility of using them to formulate arguments. From such a principle, and the premise that some object *O* is *F*, one can infer the aesthetic claim, that *O* is *G*. Principles thus promise to make room in aesthetics for something that counts as argument, in the clearest possible sense of that term.

---

[14] For other examples, see Budd, *op. cit*; Paisley Livingston, "On an Apparent Truism in Aesthetics," *British Journal of Aesthetics*, XLIII, 3 (July 2003): 260–78; and Alan Goldman, "The Experiential Account of Aesthetic Value," *Journal of Aesthetics and Art Criticism*, LXIV, 3 (Summer 2006): 333–42.

[15] As with the Requirement, it would be best to distinguish the Acquaintance Principle as an epistemic norm from its correlate governing Use by reformulating the former: "If an aesthetic belief is to count as knowledge, one must have experienced for oneself the object it concerns." That would improve upon the formulations in the literature, all of which are ambiguous between these two interpretations.

[16] Immanuel Kant, *The Critique of Judgement*, J. C. Meredith, trans. (Oxford: Clarendon, 1952), §34

Consider some sound argument taking a principle of taste as premise. Is grasping such an argument a legitimate source of aesthetic belief? The Requirement rules that it is, or at least might be. For to understand the argument might be to grasp the conditions in virtue of which the aesthetic claim forming its conclusion, that *O* is *G*, is true. I suppose it is possible that some factors universally correlating with the presence of *G*-ness might nonetheless fail to provide such conditions. If so, the existence of a sound argument from a principle of taste does not guarantee that the Requirement will be met by anyone persuaded by that argument. But, even if grasping such arguments is not always enough to meet the Requirement, surely it often will be. For often the principle will capture, not merely a necessary correlation, but one with the right sort of depth: one where *O* is *G* in virtue of the fact that it is *F*. Whenever this is so, the Requirement rules that grasping sound arguments from principles gives one the right to the resulting belief.

Contrast the Acquaintance Principle. Grasping an argument, of whatever form, is never sufficient to perceive the object its conclusion concerns. Thus understanding such an argument is never sufficient to meet the Acquaintance Principle. The two norms thus differ in their consequences. One excludes, the other allows, appeal to principles of taste in forming aesthetic belief.

The argument over principles of taste has concentrated on whether there are any.[17] It has been taken for granted that, if there are, the arguments they allow us to formulate can be used to derive aesthetic belief. Perhaps not coincidentally, discussion of the Acquaintance Principle has tended to concentrate on its implications for testimony, and ignored its implications for principles of taste.[18] Insofar as a connection has been made between the Acquaintance Principle and principles of taste, the assumption has been that the former is inconsistent with the existence of the latter. But all this assumes that the Acquaintance Principle is an epistemic norm. We now see another possibility: it might be a norm of Use. If so, the dialectic shifts considerably. So understood, the Acquaintance Principle excludes direct reliance on principles of taste to form aesthetic belief. But it does so whether or not there are any such principles. An advocate of the Principle thus need not shoulder the burden of arguing that there are no principles of taste, and an advocate of aesthetic argument needs to do more than show that such principles exist. Even if they do, it might still be illegitimate to exploit them.

---

[17] See, for example, George Dickie, "Iron, Leather and Critical Principles," in Matthew Kieran, ed., *Contemporary Debates in Aesthetics and the Philosophy of Art* (Malden, MA: Blackwell, 2005), pp. 313–26; and Goldman, "There Are No Aesthetic Principles," pp. 327–54 in the same volume.

[18] In part, this may be because the Principle is often formulated as the demand that we have "first hand" experience of the objects of aesthetic belief. Reliance on testimony is clearly "second hand," as appeal to argument is not.

VII

I have not endorsed either of the candidate norms of Use, the Requirement or the Acquaintance Principle. Perhaps one or the other governs our aesthetic thinking; perhaps some further norm of Use does so; or perhaps aesthetic thinking is not governed by norms of Use at all. However, without settling this issue we are now in a position to uncover another benefit of framing pessimism in terms of Unusability. It allows us to deal with a serious threat to any pessimist position.

The threat comes from the homely thought that we often take the recommendations of others on aesthetic matters, and that we are right to do so. Consider filmgoing again. If a friend recommends a film she has seen, surely there need be nothing wrong with my going to it on that basis. But how can this be, if, as pessimism claims, it is illegitimate to take my aesthetic views from the testimony of others? We can formulate this challenge as a little argument that brings out how it threatens both forms of pessimism:

(1) It can be perfectly licit to go to a film, rather than the others available, on a friend's recommendation.
(2) A necessary condition of this action being licit is that it be rational.
(3) Such action can only be rationalized by appeal to a warranted belief that the film is worth seeing (more so than at least most of the rivals).

So

(4) Provided the belief is true, the testimony is a source of aesthetic knowledge. (So Unavailability pessimism is false.)

And

(5) Since it is perfectly licit to act in this way, there is no further non-epistemic difficulty in acquiring aesthetic beliefs by these means. (So Unusability pessimism is false.)

What is the pessimist to say in reply? For the Unusability view, I suggest, the response is clear. Norms of Use, like many other norms, do not hold come what may. There are circumstances in which they lapse. What will those circumstances be for the norm putatively governing aesthetic thinking, be it the Requirement, the Acquaintance Principle, or some other? The effect of that norm is to require one either to think aesthetic matters through for oneself (at the least, to experience for oneself the object in question), or to remain agnostic. What it excludes is taking one's view on the matter from others. But what if neither of the options the norm permits is available? What if one can neither stay agnostic, nor settle the matter for oneself? In those conditions, I suggest, the norm will lapse. In effect, then, the application of this norm is governed by the higher-level norm that ought implies can.

This idea promises to allow Unusability pessimism to handle various putative problem cases, including those concerning aesthetic education, and those involving works now lost. I will work it through only for the case in hand. How does the idea that the norm of Use might lapse bear on our sometimes taking recommendations from others? These are taken, I suggest, when the conditions are precisely those under which the norm lapses. There are many films showing at any one time, some no doubt worth seeing, others not. Assuming I want to see a film at all, how am I to choose which to go to? If I remain agnostic about the merits of each of them, then I must either not go to the cinema at all or choose one at random. The former is perverse, given my desire to see a movie; and the latter is risky, given that the quality of what is on offer usually varies considerably. Agnosticism, then, is not a genuine option. But neither is investigating the matter for myself. That would require me to see all the films, to find out which is most worth seeing. And, while I have the time and the desire to see one film, I have neither the time nor the desire to see them all. In sum, since I can neither remain agnostic nor settle the matter for myself, the norm of Use lapses. And since Unusability views do not deny that testimony can make aesthetic knowledge available, there is nothing puzzling in how it can then be rational for me to take my friend's word that this film is decent, or that it is not.

Optimists might concede that this description of the case is coherent, but ask why it should be preferred to their own: that recommendations can be taken because, quite generally, there is nothing—epistemic or otherwise—wrong with aesthetic testimony. The Unusability view has an answer. It is that any force my friend's recommendation has for me is purely *pro tempore*. Once I have seen the film for myself, her view should count for nothing in my assessment of it. The force of this reply is not apparent if we concentrate on cases in which I form a strong opinion. Whether I hate the film or love it, the optimist will be able to finesse the fact that the recommendation now counts for very little by treating it as simply outweighed. It still counts for something, but I am bound, having seen the film, to put far greater weight on my own experience. But consider a case in which seeing the film leaves me uncertain what to think. My own experience fails to settle the matter. Can I then continue to appeal to my friend's recommendation, in forming my view? The pessimist will say not, and I am strongly inclined to agree with her. For all that the recommendation rightly had weight with me before, I must now discount it altogether. Optimism is at a loss to explain why my friend's testimony should have lost all significance in this way. If the recommendation counted epistemically before, why should it not do so now? After all, on the optimist view, the only norms operative here are epistemic. If my experience leaves the question of the film's merit in the balance, why shouldn't my friend's testimony tip it one way or the other? The Unusability view, in contrast, can easily accommodate the recommendation's shifting status. Now that I have investigated the film for myself, the norm of Use once again applies. That forbids me to appeal to anything but my own deliberations and experience in forming my view of the merits of the piece. If my experience fails to settle the matter, agnosticism is the only course permitted. The recommendation remains epistemically significant, but it is now illicit to use that resource.

Of course, the dyed-in-the-wool optimist may be unimpressed by all this. She may insist that, even once I have seen the film, I can give some weight to my friend's opinion. That seems wrong to me, but I cannot argue the determined optimist out of that view. This is one of the points at which the standoff in intuitions returns to stifle debate. The point, however, is that the Unusability view has nonetheless made a good deal of progress. For the recommendation argument began with a premise (1 above) that no one should reject. If pessimism is inconsistent with that, it really is in trouble. The effect of the above is to show that the Unusability view can readily accommodate that homely thought. If the debate then resolves onto inherently controversial matters, Unusability pessimism has deflected a serious threat, and replaced it with an open question.

What of Unavailability pessimism? Can it likewise deflect the threat posed by the argument from recommendations? That is a complex question, one that I do not have the space to go into here. My suspicion is that meeting the challenge will require the view to adopt some form of contextualism about knowledge, and thereby to complicate its theoretical framework, and to burden itself with further commitments. Perhaps Unavailability can meet the challenge; Unusability's ability to do so is proven.

VIII

I close by noting a major issue about which as yet nothing has been said. Suppose the pessimist settles on the Requirement, or the

Acquaintance Principle, or some third norm as yet unarticulated, as the norm governing aesthetic belief, and hence as the source of the illegitimacy of forming such belief by taking (pure) testimony. Why is our aesthetic practice governed by that norm? What about that practice is responsible for its having the normative structure thus described? What is the source of that norm's authority over us? (I intend these to be three formulations of the same question.)

What might an answer look like? In the moral case, we have at least some idea. If moral belief is governed by the Requirement (for morals), there are at least the materials, in existing moral theories, to explain why it might be. We might appeal to the Kantian idea, refined by Korsgaard, that first-order moral norms can only have authority over us if we understand the sources of those norms in our own legislation.[19] If subjection to a norm requires us to will it as law, and if the willed law is always an abstract principle, as Kant thought, then only she who grasps the principle underlying what morality dictates can be subject to its demands. Alternatively, we might appeal to Scanlon's thought, also ultimately deriving from Kant, that morality is essentially about being able to show to others that one's demands are rooted in considerations that transcend the personal.[20] Developing these accounts so that they meet the current need would be a major task. Nonetheless, in the moral case, we at least have some idea where to begin.

The situation in aesthetics is perhaps less promising. None of the starting points available in the moral case transfers at all naturally, and there is rather less theory already developed that might provide the source of alternative accounts. One might again hope for help from Kant, at least in the third Critique. Indeed, Kant was as pessimistic about aesthetic testimony as he was about testimony in morals. However, much of his discussion implicitly assumes that the problem is one of Availability. Since in my view nothing in his account of aesthetic judgment explains why testimony on such matters should fail to yield knowledge, it is perhaps even less likely that it will explain how that testimony is subject to a norm of Use.[21] More recently, the odd defender of the Acquaintance Principle has sought to ground it in the idea that art's value lies in the experiences it sustains.[22] Again, however, the assumption has been that the Principle is one governing

---

[19] Christine Korsgaard, *The Sources of Normativity* (New York: Cambridge, 1994).

[20] T. M. Scanlon, *What We Owe to Each Other* (Cambridge. Harvard, 1998).

[21] I explain my skepticism about the prospects for Kant's view in "Beauty and Testimony." For a very interesting discussion that instead treats Kant as an Unusability pessimist, see Keren Gorodeisky, "A New Look at Kant's View of Testimony," *British Journal of Aesthetics*, L, 1 (January 2010): 53–70.

[22] See Goldman, "The Experiential Account of Aesthetic Value."

Availability; and the attempt to ground it is more obviously riven with lacunae than was Kant's. None of this, of course, shows that attempts to ground the Principle, or other aesthetic norms of Use, are bound to fail. The very novelty of the idea of such norms suggests that there may be much remaining to be done in exploring the options here. I merely note how primitive our position is, in seeking to explain where the authority of such norms might stem from.

Of course, it might be that that question admits of no answer. Perhaps the fact that our aesthetic practice is governed by this norm is simply brute. Perhaps the only explanation in the offing is historical, of how it came about that our practice has this feature. Perhaps nothing can be done to ground that feature rationally in the deeper aims or point of aesthetic engagement.

Perhaps indeed. However, if so, the Unusability pessimist might find herself under increased pressure in another area. There is the question why the norm holds, and the question why we should believe it does. Her answer to the second can appeal to her intuitions about particular cases of the formation of aesthetic belief, and to any more theoretical reflections she has on why testimony is not a legitimate source for it—at least if those reflections are at a level below appeal to the norm of Use itself. But, as noted at the start, appeal to intuition can rapidly lead to stalemate. If our pessimist is not simply to articulate a view of the area to her own satisfaction, but also to furnish herself with grounds by which to persuade her opponents, she would do well to have an account of why the norm holds. Without that, optimists will find it easier to stick to their guns and deny that it does. And others, though persuaded that we act as if that norm governs aesthetic matters, might wonder whether it is reasonable for us to do so.

ROBERT HOPKINS

University of Sheffield

SEPARATING DIRECTIVES AND ASSERTIONS USING SIMPLE SIGNALING GAMES*

Most contemporary accounts of meaning involve, in one way or another, the intentions of the speaker and/or listener. In contrast, David Lewis develops an account of meaning in *Convention* which does not require reference to intentional states.[1] He instead defines meaning as exclusively a feature of a social group—it arises when speakers and listeners behave in mutually compatible ways in an appropriately defined game. Because intentions can be left out of this account, meaning can be present in any group which can arrange itself appropriately, including many animal species. Since many examples of animal signals appear to miss a variety of features considered central to human languages, some regard this as a reductio of Lewis's account. Others have suggested that the meaning present in animal signals (and in Lewis's toy model) represents "primitive content" which can be further refined in various ways to arrive at meaning in human language.[2] Some of these refinement accounts attempt to make intentionality a derivative notion which emerges from language (or alongside it), rather than a requirement for it.[3]

These projects which refine Lewis's definition of meaning must address features of human language that cannot be captured in Lewis's simple game. Here I focus on one shortcoming of Lewis's account, its failure to distinguish between directives and assertions.[4] I present a new Lewis-style game which accounts for this distinction and thus demonstrates how, in this case, Lewis's strategy might be pushed forward to a more complete account of human language.

---

* The author would like to thank Bill Harms, Simon Huttegger, Mandy Simons, and Brian Skyrms for their helpful comments on earlier drafts. This paper grew out of an interesting conversation with Kevin Kelly and Peter Kriss.

[1] David Lewis, *Convention: A Philosophical Study* (Cambridge: Harvard, 1969).

[2] William F. Harms, *Information and Meaning in Evolutionary Processes* (New York: Cambridge, 2004), and "Primitive Content, Translation, and the Emergence of Meaning in Animal Communication," in D. Kimbrough Oller and Ulrike Griebel, eds., *Evolution of Communication Systems: A Comparative Approach* (Cambridge: MIT, 2004), pp. 31–48, Ruth Garrett Millikan, *Language: A Biological Model* (New York: Oxford, 2005).

[3] See Millikan, *op. cit.*

[4] John R. Searle, "A Taxonomy of Illocutionary Acts," in *Expression and Meaning: Studies in the Theory of Speech Acts* (New York: Cambridge, 1979), pp. 1–29.

0022-362X/11/0803/158–169

I. LEWIS SIGNALING GAMES

Central to Lewis's account of meaning is the definition of a sort of language game, which has come to be known as a *Lewis signaling game*. The story of Paul Revere serves as a motivating example. The sexton of the Old North Church and Revere face a coordination problem. The sexton can observe whence the British troops arrive but can do nothing about it. Revere can inform the city but cannot observe the troops. Seeing that together they can achieve an outcome they both desire, they agree on a convention: the sexton will hang one lantern if the British come by land and two if they come by sea. In this circumstance, few would disagree that *one lantern* means the British are coming by land and that *two lanterns* means they are coming by sea.

Generalizing this story, Lewis describes a simple game which he believes captures an important part of linguistic behavior. In this game there are two players, one named *the sender* and one *the receiver*. The sender is aware of some aspect of the world that the receiver is not, and she has at her disposal a set of signals which she can send to the receiver, who must then take some action. The sender and the receiver both benefit if the receiver takes an action which is appropriate to the state, and the appropriate act differs from state to state.

In the game's simplest form, there are two relevant states of the world, two signals, and two actions that the receiver can take. *Action one* (A1) is appropriate for *state one* (S1), and *action two* (A2) is appropriate for *state two* (S2). The signals are labeled Sig A and Sig B. We will also suppose for simplicity that both the sender and the receiver win when the appropriate action is taken and lose when an inappropriate one is taken.

This situation describes a simple game, which has many different Nash equilibria.[5] Of particular interest are the Nash equilibria dubbed by Lewis as *signaling systems*. In the simple two-state, two-signal, two-act game there are two such equilibria. One occurs when the sender sends Sig A in S1 and Sig B in S2 and the receiver performs the appropriate acts in response to the signals. The other signaling system permutes the signals for both the sender and receiver. These equilibria are notable because they are *Pareto superior*—both players fare better in these equilibria than in any others—and *strict*—each player would do strictly worse by unilaterally deviating.

Even if we suppose that individuals will always coordinate on Pareto superior or strict equilibria (a very strong assumption), there are two choices, and there remains a question of how individuals can come to coordinate on one signaling system rather than the other. Notice that if they disagree about which signaling system is in force they will completely miscoordinate and always lose. Lewis suggests that natural salience or prior agreement (as in the case with Revere) may contribute to coordination on a signaling system.

---

[5] A set of strategies (contingency plans over all possible situations) constitutes a Nash equilibrium if and only if no player would do strictly better by changing her strategy while the other players keep their strategies fixed.

More recently, several scholars have focused on how individuals might come to play these strategies without the aid of prior agreement or salience.[6] In these models prior agreement is replaced by an evolutionary or learning process where individuals can come to coordinate on a signaling system with very weak assumptions about their underlying cognitive ability.

Lewis argues that in signaling system equilibria the signals have acquired meaning.

> I have now described the character of a case of signaling without mentioning the meaning of the signals; that two lanterns meant that the redcoats were coming by sea, or whatever. But nothing important seems to have been left unsaid, so what has been said must somehow imply that the signals have their meanings.[7]

Of course, intentions may often be involved in the meaning of signals even for Lewis. Intentions can shape the emergence of a signaling system and can be involved in its being sustained. However, one might come to these equilibria in many ways which do not involve intentional decisions at all—this is a central feature of the evolutionary accounts. These evolutionary accounts offer a sort of proof of possibility that much of human-like meaning can be had without an appeal to intentions. The evolutionary models, which describe how individuals come to play signaling systems, are akin to showing how meaning could emerge from a context where nothing is meaningful.[8]

Lewis's account opens up the possibility that nonhuman signaling systems might also be endowed with meaning. For instance, baby birds often signal their mothers or fathers of their hunger by begging. Extensive experimental work has been done, and it has been determined that in many species the chicks alter the amount or degree of begging dependent on their hunger, and parents determine how much food to give a chick based on the amount of begging.[9] As a result, we can describe a game—similar to Lewis's—where the chicks and the parent are in equilibrium.[10] As a result, a theory of meaning which only requires that the players be in equilibrium would entail that the chick's begging means that the chick is hungry.

---

[6] For a review, see Simon Huttegger and Kevin J. S Zollman, "Signaling Games: Dynamics of Evolution and Learning," in Anton Benz, ed., *Language, Games, and Evolution* (New York: Springer, 2011).

[7] Lewis, *op. cit.*, pp. 124–25.

[8] Harms, *op. cit.*; Brian Skyrms, *Evolution of the Social Contract* (New York: Cambridge, 1996), and *Signals: Evolution, Learning and Information* (New York: Oxford, 2010).

Many have found this conclusion unsettling. Human languages are of a different kind than animal languages, and meaning is uniquely a feature of our languages. As a result, scholars have attempted to point to the feature of human language which simultaneously differentiates it from animal language and which endows human language with meaning. Since many examples of human language use involve intentions, intentionality is often invoked to distinguish human from animal language. A naked appeal to intentionality is insufficient, however, since those who defend the meaningfulness of animal language are challenging exactly this claim—that meaning is essentially intentional. Instead, one must illustrate what, other than intentions, is left out of this nonintentional account of meaning. There are many potential avenues, each of which deserves its own discussion. We will here focus on the distinction between directives and assertions, which are often marked in human languages by the indicative or imperative mood.

II. ILLOCUTIONARY FORCE

Sender	Receiver
S1 → Sig A	Sig A → A1
S2 → Sig B	Sig B → A2

Table 1: A signaling system

Lewis and those after him have recognized the inadequacy of Lewis signaling games in explaining all the nuances of linguistic behavior. One can capture this shortcoming by considering how one might translate the signals of Table 1. Sig A might be translated as "we are now in S1," or alternatively as "you should do A1."[11]

These two English translations might lead one to object that meaning is not present in these signals because they do not capture a fundamental distinction. More modestly, one might concede that something like meaning is there, but understanding the emergence of this meaning is unlikely to help us understand how terms in our languages come to have the meanings they do.

---

[9] William A. Searcy and Stephen Nowicki, *The Evolution of Animal Communication* (Princeton: University Press, 2005).

[10] There is some disagreement over exactly what game is being played and exactly what the equilibrium is, but for our purposes such details can be ignored. For a discussion see *ibid.* and John Maynard Smith and David Harper, *Animal Signals* (New York: Oxford, 2003).

[11] See Harms, "Primitive Content"; Millikan, *op. cit.*

The two English translations syntactically differ—one is an indicative and the other an imperative—but the concern can also be captured by considering the different illocutionary forces which this syntactic distinction usually marks. The sexton might want a single lantern to simply inform Revere about the British, or he might want a single lantern to cause Revere to ride to Lexington yelling "The British are coming by land!" The former has the force of an assertion, transmitting information, while the latter has the force of a directive, encouraging another to act in a certain way. Searle describes this difference as a difference in "direction of fit."[12] Assertions are meant to fit some preexisting fact in the world, while directives are meant to cause the world to fit the sentence. Searle characterizes this distinction in a thoroughgoing intentional way, but in keeping with the Lewisian strategy we might see if this can be done away with.

Rather than engage intentional meaning, Brian Skyrms suggests that a quantitative notion of information might capture all that is needed to explicate the meaning of the signals in a Lewis signaling game.[13] Consider the signaling-system equilibrium pictured in Table 1. Suppose that prior to hearing a signal or seeing an action the probability of S1 is 0.5. In the signaling system of Table 1, the conditional probability of S1 given that the sender sent Sig A is 1.0. In this sense, the signal carries perfect information about the state. But there is also perfect information about the act of the receiver. Conditioned on the sender sending Sig A, the probability that the receiver will take A1 is also 1.0. Here, a sort of symmetry parallels the two English translations of Sig A: it carries perfect information about both the signal and the act. The central argument of this paper is that, in order to account for the distinction between directives and assertions, we must find a way to break that symmetry, that is, to find a game where (in equilibrium) the signal carries information about the state and not the act or, alternatively, about the act and not the state. If we do this in the right way, we might be able to find the nonintentional equivalent of Searle's "direction of fit."[14]

---

[12] Searle, *op. cit.*

[13] See Skyrms, *Signals*. This is not to say that a quantitative notion of information can always capture the representational content of signals in every context, just in the case of Lewis signaling games. Peter Godfrey-Smith, for instance, has suggested that a notion of information cannot provide a naturalistic foundation for indication generally. See Godfrey-Smith, "Indication and Adaptation," *Synthese*, xcii, 2 (August 1992): 283–312.

III. HUTTEGGER'S DISTINCTION

Following a suggestion by Lewis and Harms,[15] Simon Huttegger attempts to model the syntactic difference between indicatives and imperatives as a difference in deliberation.[16] He describes a signaling game where, in addition to the normal acts, both the sender and the receiver must choose whether or not to "deliberate."

Sender	Receiver
S1 → (Do not deliberate, Sig 1)	Sig 1 → (Deliberate, A1)
S2 → (Deliberate, Sig 2)	Sig 2 → (Do not deliberate, A2)

Table 2: Huttegger's signaling system

He considers a situation where one of the players must deliberate in one state and the other must deliberate in the other state. Table 2 shows an example of a signaling system in this game—in one state only the sender deliberates, and in the other only the receiver does. Huttegger suggests that in this equilibrium we should interpret Sig 1 as indicative (or an assertion) and Sig 2 as an imperative (or a directive). This corresponds to one feature of indicatives and imperatives. I inform you that a car is coming when I believe that you need to combine this information with some of your background knowledge or desires in order to determine the correct course of action. I yell "Stop!" when I believe I know the correct course of action, and I believe that any delay on your part would be harmful.

"Deliberation," however, is just a name attributed to some additional action added in Huttegger's model. The appropriate act in S1 is for the receiver to deliberate and then do A1. There is no combination of information or choice of additional action in either case.

---

[14] Not every way of breaking this symmetry would be sufficient to capture this distinction Harms (in conversation) gave an insightful counterexample. Suppose that the sender is perfect in her implementation of a signaling system, but the receiver often makes mistakes. There will be an asymmetry in information because the signal carries perfect information about the state but less than perfect information about the act, but it would be odd to call the signal assertive as a result. I take it as necessary, if not sufficient, to break this symmetry in order to account for the difference between assertions and directives In the case I present below, I break it in one extreme way—by giving either perfect or no information about the act or state. However, symmetry might also be broken in less extreme ways.

[15] Lewis, *op. cit*; Harms, *op. cit.*

[16] Huttegger, "Evolutionary Explanations of Indicatives and Imperatives," *Erkenntnis*, LXVI, 3 (May 2007): 409–36.

Huttegger leaves this process at a more abstract level, supposing that a refined model could be constructed with similar results.

Although he describes his model as developing a distinction between indicatives and imperatives, Huttegger's model does capture one asymmetry between directives and assertions. Often we direct so that others need not think about what to do, and we assert so that others might decide the best action to take. However, one might object to Huttegger's model by raising the same problem as before. It seems as though signals in Huttegger's model retain the same two translations that were present in Lewis's original game. For instance, Sig A might be translated both as "deliberate and then do A1" and "we are in S1." Huttegger suggests that the latter is a more natural interpretation, but of course the former remains available.

This point can be underscored by appealing to the quantitative notion of information discussed above. As with the original Lewis signaling game, the signal gives perfect information both about the act and about the state. In the signaling system pictured in Table 2, the probability of A1 is 1.0 given the receipt of Sig A, even though this signal is the supposed indicative/assertive signal.

One might extend Huttegger's model in various ways to secure this distinction. For example, one might break A1 into two actions (say, A1.1 and A1.2). The deliberation by the receiver is necessary to determine which of these two actions is correct. Similarly, one might do the same for action A2 but require the sender to deliberate and then send one of two signals. Here, something like informational asymmetry might be constructed in Huttegger's models. As I will explain in the next section, while this is possible, I believe we can adequately model directives and assertions without any appeal to deliberation whatsoever.

IV. A NEW GAME

In order to illustrate the alternative game, I will construct a Lewis signaling game with one extra receiver. Although he has not considered a game like the one presented here, Skyrms has begun to study these types of modifications and the implications they have for the evolution of meaningful signaling systems.[17]

Consider the awkward situation of moving a piece of furniture, like a couch, that requires two people to carry. Often it is helpful to have a third person in position to observe the orientation of the object and direct the movement. That third person might help to level the couch either by describing the position of the couch—"the back is too low"—or by directing the actions of the movers—"you on the back, lift up." In the former case, the director often can issue only one statement, leaving it to the movers to position themselves; however, in the latter he must issue two: a directive for each mover. The director would assert in the former case and would direct in the latter.

---

[17] Skyrms, *Signals*, and "Evolution of Signaling Systems with Multiple Senders and Receivers," *Philosophical Transactions of the Royal Society of London B*, CCCLXIV (2009): 771–79.

We will abstract away from cases like this in much the way Lewis did. In our game there will be one sender who is aware of the state of the world and two receivers that must take complementary actions. All three are rewarded only when both receivers take the appropriate actions. Suppose for simplicity that there are only two states, although this assumption is not critical to the discussion that follows. Suppose that the appropriate actions for each receiver are as listed in Table 3.

		States	
		S1	S2
Receivers	R1	A1	A2
	R2	A2	A1

Table 3: Appropriate actions for each state

The receivers must take complementary actions, and the state determines who should do what. In our furniture-moving example, we might imagine that the two states represent the tilt of the couch (either forward or backward). If the couch is tipped forward, the front mover must lift up while the back mover must lower, and vice versa when the couch is tipped backward.

If we allow the sender to differentiate the message that she sends to each receiver, we have a total of four signaling-system equilibria (there are many other equilibria as well). I will call two of these signaling-system equilibria *a-equilibria*. In the a-equilibria the sender sends the same message to both receivers using a different signal in different states. One such equilibrium is pictured in Table 4.

Sender	Receiver 1	Receiver 2
S1 → Sig A (to R1) S1 → Sig A (to R2)	Sig A → A1	Sig A → A2
S2 → Sig B (to R1) S2 → Sig B (to R2)	Sig B → A2	Sig B → A1

Table 4: An a-equilibrium

Consider the informational content of each signal. Like the simple Lewis signaling game discussed above, Sig A (to either receiver) carries perfect information about the state. What about the act? If we restrict ourselves to considering only one particular receiver (say, Receiver 1), then here too the signal carries perfect information about the act. But if we average over both receivers we find that this information is lost. The probability that a randomly chosen receiver will take A1 given that she is sent Sig A is equal to the unconditional probability that the receiver will take A1. Here we have one example of the informational asymmetry we sought. The asymmetry suggests that we have an assertion because it carries perfect information about the state but no information about the act.

We can generate directives as well. Consider the signaling system pictured in Table 5, which I will call a *d-equilibrium*.

Sender	Receiver 1	Receiver 2
S1 → Sig A (to R1) S1 → Sig B (to R2)	Sig A → A1	Sig A → A1
S2 → Sig B (to R1) S2 → Sig A (to R2)	Sig B → A2	Sig B → A2

Table 5: A d-equilibrium

Here, in both states the sender sends both signals to a receiver but conditions to whom he sends each signal on the state. In this equilibrium it appears as though the signal means something about what the appropriate act is but says nothing about the state. Again, we can consider the informational content of each signal. Globally, there is now perfect information about the act, since the probability that a randomly chosen receiver takes A1 given he received Sig A is 1.0. There is local information about the states here, too. Given that Receiver 1 receives Sig A, the probability that we are in S1 is 1.0. However, if we average over both receivers, that information is lost. Knowing that a randomly chosen receiver has received Sig A gives us no information about the state, despite the fact that it gives us perfect information about the act. Again we have an asymmetry, this time one that suggests the signals are directives.

We have developed an informational asymmetry that corresponds to the distinction between assertions and directives. In a-equilibria there is perfect global information about the state but not the act—the signals are assertions. In d-equilibria there is perfect global information about the act but not the state—the signals are directives.

One might object to this model by noting that it appears to require three-person conversations in order to get off the ground. The distinction between directives and assertions is perfectly coherent in two-person conversations as well. How, then, is this model to account for that? The three-way conversation described here is merely a heuristic; the game is equivalent to many two-person conversations. Suppose there is just one sender and one receiver. The sender observes the state of the world while the receiver remains ignorant, just as before. But now suppose that the receiver is in one of two "positions," namely, either Receiver 1 or Receiver 2, and suppose that both the sender and receiver can observe her position. The actions and payoffs are just as before, but now there is a single right action which depends on the state and the position of the receiver.

This two-person game is essentially equivalent to the game described above; it has all the same equilibria (as well as the same strategies and payoffs). Now imagine a population of players all observing the same linguistic conventions but having different relevant states and positions. Averaging over that population in an a-equilibrium will result in the signal carrying perfect information about the state and zero information about the action taken by the receiver, and averaging over a population in a d-equilibrium will reverse the role of the signal.

This two-player version suggests other possibilities as well. What if the receiver's position was not known by both players? If the receiver knows her own position but the sender does not, there cannot be any d-equilibria because the relevant strategies are not present.[18] The language here can only have assertions. Given that the sender does not know what the correct action is, how could he issue a directive? This is the situation faced by Cambell's monkeys, who give an alarm call when they spot a predator.[19] Since the right response for a listener depends on the situation that listener is in—and since the signaler is signaling to many different monkeys in different contexts of which the signaler is unaware—it seems proper to interpret the signal as an assertion. We might also imagine another situation where only the sender is aware of the receiver's position, and here only the d-equilibria would be possible. Again this seems plausible; if I have sufficient information to know the right action but you do not, I should issue a directive. Finally, this model might be expanded to consider situations where the position itself is a signal sent by another signaler, by nature, or by the perceptive apparatus of the receiver. Modifications of this sort would be needed to fully account for the evolution of such terms, but I will not consider them in detail here.

[18] I am indebted to Brian Skyrms for this suggestion.

[19] Klaus Zuberbühler, "Predator-Specific Alarm Calls in Campbell's Monkeys, *Cercopithecus campbelli*," *Behavioral Ecology and Sociobiology*, L, 5 (2001): 414–22

While this model overcomes some of the challenges posed to Huttegger's model, I do not mean to say that this model is right and that Huttegger's is wrong. In fact, features of both models may be present in actual situations of indicative/assertion and imperative/directive use. The point here is not so much to illustrate the one true model, if there is such a thing, but instead to illustrate the strength and flexibility of this approach for developing models of many different linguistic phenomena.

V. THE INTENTIONALITY CHALLENGE

With respect to the directive/assertion distinction, has anything important been left unsaid? Does this game-theoretic account capture everything important about the behavior? No doubt many will regard it as inadequate. The directive/assertion distinction is usually characterized with ample references to the intentions of the speaker (and perhaps the hearer). My account does not require the speaker or the hearer to be the sort of thing that has intentions at all.[20]

One cannot now just require that any adequate account of this distinction make reference to intentions. The whole aim of the Lewisian project (or at least my characterization of it) is a minimalistic one—to see how much of linguistic behavior can be captured without appeal to intentions at all. If my account has failed to capture something central about the directive/assertion distinction, that gap must be explicated in nonintentional terms for the objection to not beg the question.

Following this line of argument, one might continue to push the objection from translation discussed above. In the a-equilibrium pictured above, one can interpret Sig A to mean "do A1 if you are Receiver 1, and do A2 if you are Receiver 2." Similarly, in the d-equilibrium one can interpret Sig A to mean "we are in a state where the best action for you is A1." We seem to have returned to the problem that was initially posed for Lewis and again for Huttegger.

It seems clear that in this case the plausibility of the two translations is stretched further than it was before, but not that those translations are somehow impossible. However, those who would point out that we have not eliminated the possibility of devising both indicative (or assertive) and imperative (or directive) translations of Sig A might have difficulty describing what would be necessary to demonstrate that such a real distinction exists. Even English indicatives and imperatives often convey some information about the state and some information about the act. Knowing that I have said "run" to someone gives you some information about the world, namely, that the state was such that I wanted my conversant to run.

---

[20] As in Lewis's model, intentions *can* play a role in explaining how people come to play particular equilibria, but it *need not* play a role.

I do not believe that the possibility of these two translations demonstrates an inadequacy of the Lewis signaling games as a model for linguistic behavior. Instead, this possibility is an example of the more radical problem of translation suggested by Quine.[21] Merely because I can offer two different English sentences which both capture the meaning of the signals in the game is not sufficient to demonstrate that meaning is not present in these games, just as the presence of two possible interpretations of the term 'gavagai' does not demonstrate that the natives are not using meaningful language.

The onus instead should be on the objector to point out what exactly is missing from these games. The challenger must provide some test which one can apply to another human language in order to determine if that language has a directive/assertion distinction that can also be applied to artificial settings like signaling games. Merely asserting that something is not present in these cases leaves open the possibility that French (or Japanese, or whatever) may too be missing that feature, a conclusion that I believe no one would regard as appropriate.

KEVIN J. S. ZOLLMAN

Carnegie Mellon University

---

[21] W. V. O. Quine, *Word and Object* (Cambridge: MIT, 1960).

CAMBRIDGE

NEW *and* EXCITING TITLES

Ethics and Animals
An Introduction
LORI GRUEN
Cambridge Applied Ethics
$85 00 Hb 978-0-521-88899-8
$27 99 Pb 978-0-521-71773-1
250 pp

Kant: Observations on the Feeling of the Beautiful and Sublime and Other Writings
PATRICK FRIERSON *and* PAUL GUYER
Cambridge Texts in the History of Philosophy
$99 00 Hb 978-0-521-88412-9
$28.99. Pb 978-0-521-71113-5
396 pp

Now in Paperback!
Anthropology, History, and Education
IMMANUEL KANT, ROBERT B LOUDEN *and* GÜNTER ZÖLLER
The Cambridge Edition of the Works of Immanuel Kant in Translation
$49 00 Pb 978-0-521-18121-1
616 pp

Immanuel Kant: Groundwork of the Metaphysics of Morals
A German-English Edition
IMMANUEL KANT, MARY GREGOR *and* JENS TIMMERMANN
$85 00 Hb 978-0-521-51457-6
212 pp

Perfecting Virtue
New Essays on Kantian Ethics and Virtue Ethics
LAWRENCE JOST *and* JULIAN WUERTH
$90.00 Hb 978-0-521-51525-2
322 pp

The Cambridge Introduction to Emmanuel Levinas
MICHAEL L MORGAN
$90 00 Hb 978-0-521-19302-3
$24 99. Pb 978-0-521-14106-2
270 pp

Immigration and the Constraints of Justice
Between Open Borders and Absolute Sovereignty
RYAN PEVNICK
$82 00 Hb 978-0-521-76898-6
210 pp

Spinoza on Human Freedom
Reason, Autonomy and the Good Life
MATTHEW J KISNER
$85 00. Hb. 978-0-521-19888-2
274 pp

Familiar Objects and their Shadows
CRAWFORD L ELDER
Cambridge Studies in Philosophy
$85 00 Hb 978-1-107-00323-1
222 pp.

Aristotle on Time
A Study of the Physics
TONY ROARK
$82 00 Hb 978-1-107-00262-3
248 pp

Socrates and Philosophy in the Dialogues of Plato
SANDRA PETERSON
$90 00 Hb 978-0-521-19061-9
310 pp

Now in Paperback!
Aristotle's Ethics as First Philosophy
CLAUDIA BARACCHI
$36 99 Pb 978-1-107-40051-1
352 pp.

The Ancient Quarrel Between Philosophy and Poetry
RAYMOND BARFIELD
$90 00 Hb 978-1-107-00032-2
288 pp

Saul Kripke
ALAN BERGER
$95 00 Hb 978-0-521-85826-7
$36 99 Pb 978-0-521-67498-0
374 pp

Heidegger, Art, and Postmodernity
IAIN D THOMSON
$90 00 Hb 978-1-107-00150-3
$27 99. Pb. 978-0-521-17249-3
256 pp

Interpreting Heidegger
Critical Essays
DANIEL O DAHLSTROM
$90.00 Hb 978-0-521-76494-0
320 pp

Paleontology
A Philosophical Introduction
DEREK TURNER
Cambridge Introductions to Philosophy and Biology
$90 00 Hb 978-0-521-11637-4
$28 99 Pb. 978-0-521-13332-6
240 pp

Conscientious Objection in Health Care
An Ethical Analysis
MARK R WICCLAIR
$90 00 Hb 978-0-521-51431-6
$29 99 Pb 978-0-521-73543-8
264 pp

Prices subject to change

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

from CAMBRIDGE UNIVERSITY PRESS

G. E. Moore: Early Philosophical Writings
THOMAS BALDWIN *and* CONSUELO PRETI
$110.00 Hb 978-0-521-19014-5
338 pp

The Ethics of Global Climate Change
DENIS G. ARNOLD
$90.00 Hb 978-1-107-00069-8
352 pp

Philosophy and Temporality from Kant to Critical Theory
ESPEN HAMMER
Modern European Philosophy
$90.00 Hb 978-1-107-00500-6
270 pp

Moral Dilemmas in Medieval Thought
From Gratian to Aquinas
M. V. DOUGHERTY
$90.00 Hb 978-1-107-00707-9
236 pp

Peirce and the Threat of Nominalism
PAUL FORSTER
$82.00 Hb 978-0-521-11899-6
272 pp

Perception and Knowledge
A Phenomenological Account
WALTER HOPP
$90.00 Hb 978-1-107-00316-3
258 pp

Elements of Moral Cognition
Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment
JOHN MIKHAIL
$90.00 Hb 978-0-521-85578-5
408 pp

Reasons for Belief
ANDREW REISNER *and* ASBJØRN STEGLICH-PETERSEN
$90.00 Hb 978-1-107-00687-4
284 pp

Herder on Humanity and Cultural Difference
Enlightened Relativism
SONIA SIKKA
$85.00 Hb 978-1-107-00410-8
292 pp

Children's Understanding of Death
From Biological to Religious Conceptions
VICTORIA TALWAR, PAUL HARRIS *and* MICHAEL SCHLEIFER
$90.00 Hb 978-0-521-19459-4
208 pp

What Literature Teaches Us about Emotion
PATRICK COLM HOGAN
Studies in Emotion and Social Interaction
$85.00 Hb 978-1-107-00288-3
352 pp

Deep Beauty
Understanding the Quantum World through Mathematical Innovation
HANS HALVORSON
$99.00 Hb 978-1-107-00570-9
486 pp

The Problem of Harm in World Politics
Theoretical Investigations
ANDREW LINKLATER
$99.00 Hb 978-1-107-00443-6
$29.99 Pb 978-0-521-17984-3
320 pp

Now in Paperback!

The Educated Brain
Essays in Neuroeducation
ANTONIO M. BATTRO, KURT W. FISCHER *and* PIERRE J. LÉNA
$37.99 Pb 978-0-521-18189-1
280 pp

Justice and Self-Interest
Two Fundamental Motives
MELVIN J. LERNER *and* SUSAN CLAYTON
$85.00 Hb 978-1-107-00233-3
280 pp

Deep Learning
How the Mind Overrides Experience
STELLAN OHLSSON
$120.00 Hb 978-0-521-83568-8
540 pp

A History of Communications
Media and Society from the Evolution of Speech to the Internet
MARSHALL T. POE
$85.00 Hb 978-1-107-00435-1
$24.99 Pb 978-0-521-17944-7
350 pp.

Infinity
New Research Frontiers
MICHAEL HELLER *and* W. HUGH WOODIN
$99.00 Hb 978-1-107-00387-3
326 pp

Prices subject to change

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

New from Oxford

The Oxford Handbook of Philosophy in Early Modern Europe
Edited by DESMOND M. CLARKE and CATHERINE WILSON
A team of leading scholars survey the development of philosophy from the mid-16th century to the early 18th century
(Oxford Handbooks)
2011 560 pp. $150.00

Figuratively Speaking
Revised Edition
ROBERT J. FOGELIN
In this revised edition, Fogelin responds to the criticisms of the first edition, updates his material, and includes a new preface
2011 144 pp. Paperback $29.95

What Should I Do?
Philosophers on the Good, the Bad, and the Puzzling
Edited by ALEXANDER GEORGE
The very best questions and answers from AskPhilosophers.org—from everyday ethical conundrums to profound questions of morality
2011 224 pp. Paperback $15.95

The Oxford Handbook of the History of Political Philosophy
Edited by GEORGE KLOSKO
Fifty distinguished contributors survey the entire history of political philosophy
(Oxford Handbooks)
2011 688 pp. $150.00

Defending the Axioms
On the Philosophical Foundations of Set Theory
PENELOPE MADDY
The latest work by an award-winning philosopher of science
2011 144 pp. $45.00

The Metaphysics of the Incarnation
Edited by ANNA MARMADORO and JONATHAN HILL
A comprehensive and in-depth overview of the metaphysics of the incarnation
2011 288 pp. $125.00

Hegel's Conscience
DEAN MOYAR
This book provides a new interpretation of the ethical theory of G.W.F. Hegel
2011 288 pp. $74.00

Consciousness and the Prospects of Physicalism
DEREK PEREBOOM
Derek Pereboom explores how physicalism might best be formulated and defended against the best anti-physicalist arguments
(Philosophy of Mind Series)
2011 240 pp. $65.00

Motive and Rightness
STEVEN SVERDLIK
An original synthesis of ethical theory, moral psychology, and philosophy of action.
2011 224 pp. $55.00

New in Paperback

Imposing Values
Liberalism and Regulation
N. SCOTT ARNOLD
(Oxford Political Philosophy)
2011 504 pp. Paperback $35.00

The Oxford Handbook of Philosophy of Mind
Edited by BRIAN MCLAUGHLIN, ANSGAR BECKERMANN, and SVEN WALTER
2011 832 pp. Paperback $55.00

Wittgenstein's Notes on Logic
MICHAEL POTTER
2011 336 pp. Paperback $35.00

Reflective Knowledge
Apt Belief and Reflective Knowledge
Volume II
ERNEST SOSA
2011 272 pp. Paperback $24.95

OXFORD

THE JOURNAL OF
PHILOSOPHY

VOLUME CVIII, NUMBER 4
APRIL 2011

page

171 *Game Theory, Indirect Modeling, and the Origin of Morality* — Arnon Levy
188 *A Mosquito Bite against the Enactive Approach to Bodily Experiences* — Frédérique de Vignemont

COMMENTS AND CRITICISM

205 *Methodology and the Nature of Knowing How* — Michael Devitt

BOOK REVIEWS

219 Debra Satz: *Why Some Things Should Not Be for Sale: The Moral Limits of Markets* — David Schmidtz
224 NEW BOOKS: ANTHOLOGIES

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2011 by the Journal of Philosophy, Inc. ISSN 0022-362X

UNIVERSITY CALCUTTA
ALIPORE CAMPUS

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 4, APRIL 2011

GAME THEORY, INDIRECT MODELING, AND THE ORIGIN OF MORALITY*

The idea that the origin of moral norms can be explained in Darwinian terms has been with us since Darwin himself. Recently, it has taken a promising form with the advent of evolutionary game theory. Combining concepts and formal tools from evolutionary biology and economics, current evolutionary game-theoretic models embody a distinct picture of the origin of morality. On this picture, morality is the product of cultural evolution. That is, the evolution of norms happens within society as the aggregate product of individual-level behavioral learning. The circumstances in which moral norms come into play are represented as strategic interactions such as bargaining and cooperation/defection scenarios. A basic assumption is that social agents generally seek to improve their lot in such interactions and that they do so by learning from peers. Strategies that have secured higher payoffs will tend to spread, because social learners will tend to adopt them more readily. The result is a Darwinian process operating on top of the basic set of human cognitive skills and motivations. Recent modeling has shown that such a process can give rise to conduct akin to cooperation, sharing, trust, and commitment.

Successes in evolutionary game-theoretic modeling have blown wind in the sails of an optimistic view according to which, as Brian Skyrms puts it, we have "the beginning of an explanation" of the origins of morality.[1] Among modelers as well as commentators, there appears to be a sense that substantial progress has been made in

---

* Many thanks to Peter Godfrey-Smith for reading several earlier drafts of this article. For comments and assistance I am also indebted to Arthur Applbaum, David Haig, Ned Hall, Frances Kamm, Daniel Rosenbloom, Susanna Siegel, and the participants at a colloquium in the Department for History and Philosophy of Science at the Hebrew University of Jerusalem.

[1] Brian Skyrms, "Sex and Justice," this JOURNAL, XCI, 6 (June 1994): 305–20, at p. 320.

0022-362X/11/0804/171–187

understanding where moral norms come from, how they change over time, and where, in principle, they might be going. But I will argue that while progress in modeling has been made, the sense that we are closer to an explanation of the actual origin of moral norms is mistaken, and the reasons underlying this mistake are worth our attention. Models of the evolution of morality contain a number of idealizations, especially with regard to how they portray individual-level moral learning. Assumptions of this sort were made early in the project and follow-up work retains them, sometimes without noticing their effect on the explanatory power of the models. Taking a close look at the idealizing assumptions involved can teach us something important, I will argue, about the applicability of Darwinian ideas to culture.

Moreover, the case of evolutionary game theory reveals an error characteristic of a certain mode of modeling. Specifically, it will be helpful to think of the models in question as employing a strategy of indirect representation and analysis. Indirect analysis explores a phenomenon in a once-removed fashion: to learn about a target system in the world one analyzes a constructed scenario ("the model"), relying on similarities between the model and the target to gain knowledge of the target from an analysis of the self-standing model. I will argue that work in evolutionary modeling of morality has largely provided us with an understanding of models, but some of the crucial similarity relations are missing. Thus, what has been achieved is more in the nature of understanding a stipulative picture of how moral change might occur in a world rather different from ours. It is not, importantly, knowledge about the actual dynamics of norms. I will offer a diagnosis of the source of the conflation of such internal progress, as I shall call it, with genuine explanatory success.

For the sake of concreteness, the discussion will focus on a family of models exploring the game "Divide the Cake." Skyrms and several collaborators have developed these models over the last 15 years or so in order to capture the origin of distributive justice.[2] The case of distributive justice is a central one which deserves attention in its own right. Since much of what I have to say about it generalizes in fairly immediate ways, I will not look at other examples. To forestall potential concerns about the kind of discussion with which I am engaged, let me emphasize that what follows does not address normative questions. The models I discuss target the processes which shape moral attitudes and moral conduct. This is a purely descriptive matter. The issue of whether an evolutionary account affects the norms we ought to hold or their epistemic standing is separate, and I do not address it here.[3] I begin, in the next section, by describing the work of Skyrms and his collaborators, paying special attention to the diachronic aspect of its development. There follows a discussion of the explanatory merits of Divide the Cake models, emphasizing the significance of some key mismatches between the models and reality. The final part of the paper is devoted to the issue of progress in indirect modeling.

---

[2] I will focus on the work of Skyrms and Jason Alexander, discussed below. Related work includes Ken Binmore, *Game Theory and the Social Contract, Vol. 1: Playing Fair* (Cambridge: MIT, 1994); Zachary Ernst, "Explaining the Social Contract," *British Journal for the Philosophy of Science*, LII, 1 (2001): 1–24, and "Robustness and Conceptual Analysis in Evolutionary Game Theory," *Philosophy of Science*, LXXII, 5 (December 2005): 1187–96; H. Payton Young, *Individual Strategy and Social Structure* (Princeton: University Press, 1998); Kevin Zollman, "Explaining Fairness in Complex Environments," *Politics, Philosophy and Economics*, VII, 1 (February 2008): 81–97.

I. THE INITIAL MODEL

The initial model, as I will call it, appeared in Skyrms's elegant 1994 paper "Sex and Justice."[4] There he proposed an account of the origin of a share-and-share-alike rule of distributive justice. Skyrms's point of departure was an apparent puzzle that arises in the context of Divide the Cake. In this game, two individuals face the problem of dividing a unit good (the cake). The allocation procedure is the following. Both players simultaneously present a demand corresponding to their desired portion of the cake. If the demands sum to one or less, each gets what she demands; otherwise, both get nothing. A salient solution to this problem is an equal 50-50 split. Game theorists as well as laymen feel an intuitive pull towards equal splitting. In experimental tests in which subjects play this game, it is by far the most common outcome.[5] When asked, subjects typically explain that equal splitting seems to be the just (or fair) way to divide the cake. However, from the point of view of rational decision making there is nothing special about a 50-50 split. It is a Nash equilibrium of the game; that is, neither player can benefit from unilaterally changing her strategy. But this holds for splitting 70-30 or 80-20 as well. The equal split is maximally efficient—none of the cake goes to waste—but so are many less equitable division schemes.

---

[3] Within the descriptive project, a more specific set of questions concerns the explanandum of an evolutionary explanation. What, exactly, is the explanatory target: is it the origin of moral intuitions? Of evaluative tendencies? Or perhaps of conduct that merely conforms with moral standards? I do not have the space to adequately discuss this issue here. I will assume, in a charitable spirit, that whatever evolutionary game theory aims to explain is closely connected with morality, and worth explaining.

[4] Skyrms, *op. cit.*

[5] John H. Kagel and Alvin E. Roth, *Handbook of Experimental Economics* (Princeton: University Press, 1998).

Skyrms proposed to look at the matter not as a decision-making problem, but as the outcome of a Darwinian process. He offered the following model.[6] Suppose a large population of agents repeatedly plays Divide the Cake. With a cake that can be sliced into *n* equal portions, the composition of the population is given by the vector $\bar{x} = x_1, x_2, x_3, \ldots x_n$ where x_i is the proportion of players demanding *i* slices of cake. Over time, the frequency of types in the population changes in accordance with the so-called replicator dynamic:[7]

$$x_i' = x_i \frac{f_i(\bar{x})}{\bar{f}(\bar{x})}$$

Here, f_i is the payoff of those demanding *i* slices—we can call them the "*i*-strategists"—and $\bar{f}$ is the average payoff in the population.[8] The replicator dynamic is one of the most commonly used equations of change in evolutionary modeling. Its key feature is that the proportion of a strategy in the population is a direct function of how it performs relative to the average payoff in the population. As f_i increases relative to $\bar{f}$, so does the growth of x_i: the more successful a strategy is, the faster it spreads. The replicator equation originated in evolutionary biology, but under certain assumptions it can be interpreted as representing a cultural dynamic driven by payoff-based imitation. If players imitate successful strategists more often than unsuccessful ones, the result is a Darwinian dynamic in which the successful strategies spread through the population.

Analyzing this model, Skyrms's main finding was that what made the difference was how players were paired. In its simplest form, the model posits a well-mixed population where players pair up completely at random. In this case an equal split can evolve, but unequal splits are quite likely too. The population may get "trapped" in an unjust state in which some players demand (and receive) more, others less. One can, however, modify the model so that pairing is nonrandom—specifically, so that an *i*-strategist has a higher-than-chance likelihood of playing against a fellow *i*-strategist. In this case we say that interactions are (positively) correlated. So modified, the model results in "demand half"[9] invariably taking over the population, and the equal split predominates.

---

[6] Young had previously offered a bargaining model for the evolution of the equal split, focusing on the same game. See "An Evolutionary Model of Bargaining," *Journal of Economic Theory*, LIX, 1 (February 1993): 149–68. Young's model is similar in spirit to Skyrms's, but some of its technical aspects are more complex and it is not as overtly designed to account for the evolution of justice.

[7] There are two differences between the initial model and a later model by Skyrms and Jason Alexander, which I discuss below. First, the later one allows slicing the cake in ten, whereas the initial model employed a more restricted strategy space. Second, the initial model used a continuous-time version of the replicator equation, rather than the discrete-time version used in the later work (and presented here) Neither point affects my arguments. For a comprehensive discussion of the replicator equation, see Jörgen W. Weibull, *Evolutionary Game Theory* (Cambridge: MIT, 1995).

[8] The vector $\bar{x}$ figures into the equation because the payoff gained by a strategist will depend on whom he or she faces.

II. ELABORATING THE PICTURE

The initial model was elegant and the analysis yielded an interesting result, but it was also a highly abstract representation of the social process in question. Agents and their strategies were represented statistically, and so were the overall effects of social learning. Correlated interaction, the key ingredient assuring reliable emergence of the equal split, was posited, but no mechanism for its generation was incorporated into the model. As a consequence, some critics challenged the rationale for Skyrms's stipulations, whereas others alleged that once more details were included, the results would not hold up.[10] Partly as a response to these criticisms, the years since the initial model appeared have seen Skyrms and a number of collaborators seeking to provide more concrete follow-ups that flesh out the initial model.

To do so, they turned to agent-based modeling. An agent-based model represents players explicitly—as individuals with a strategy, a learning rule, a record of success in the game, and often as situated within a social network. In 1999, Skyrms and Jason Alexander proposed a model of this kind for Divide the Cake, as an extension of the initial model.[11] They considered a population of 10,000 agents playing Divide the Cake on a large square lattice. In every iteration of the game each player is paired with her eight immediate neighbors in turn. The players make demands according to their strategies and garner payoff. At the end of a round, each player compares herself to her neighbors and adopts the strategy of the neighbor with the highest overall payoff for that round.[12] This way of updating one's strategy is known as "imitate your best neighbor."

Skyrms and Alexander performed computer simulations of this process, and the results were consonant with the main message of the initial model. They found that for upwards of 99% of the population configurations they simulated, the outcome was an equal split. All that was required was the presence of a small cluster of interacting "demand five"-ers—these went on to "infect" the rest of the lattice. Recall that in the initial replicator model the key requirement for the evolution of the equal split was (positively) correlated pairing. In the agent-based model this is, in an important sense, still true. Positive correlation arises from the network structure, as "new" equal-splitters are inserted near the neighbors whom they take after. Recall also that the initial model involved a payoff-driven dynamic generated by imitation learning. This too is true in the later model, where imitation is given a concrete "imitate your best neighbor" form.

---

[9] Or, correspondingly, "demand five" when the cake is sliced into ten parts, as in the models discussed below.

[10] Examples include Justin D'Arms, Robert Batterman, and Krzyzstof Górny, "Game Theoretic Explanations and the Evolution of Justice," *Philosophy of Science*, LXV, 1 (March 1998): 76–102; Neil Tennant, "Sex and the Evolution of Fair Dealing," *Philosophy of Science*, LXVI, 3 (September 1999): 391–414.

[11] Alexander and Skyrms, "Bargaining with Neighbors: Is Justice Contagious?" this JOURNAL, XCVI, 11 (November 1999): 588–98.

[12] So long as the neighbor is more successful than she is. Ties are broken by a coin toss.

* * *

In recent work, Jason Alexander (now McKenzie Alexander) has delved deeper into agent-based models of Divide the Cake.[13] In a book-length treatment of the origin of morality, he explores a range of network models, including "small world" networks, where phenomena like "six degrees of separation" are common, and bounded-degree networks, which have a constrained random structure. McKenzie Alexander also enlarges the class of imitation rules to include, for example, "imitate with proportion relative to success," or a rule that has individuals play the game with their neighbors but imitate the agent with the best score in the entire population ("imitate the winner"). In these models too there is a remarkable tendency for populations to stabilize at an equal split. As in earlier models, the combination of structure-generated correlation and payoff-based learning ensures that the demand-five strategy is superior to more (or less) demanding alternatives.

A few words about the developmental trajectory of this work will serve to summarize the discussion so far and set the stage for the argument to follow. The move from the initial replicator-based model to current agent-based models primarily involved a process of concretization: filling in specifics where the initial model contained abstract posits. The early model presupposed a payoff-driven populational dynamic, supervening on individual-level imitation learning. But it provided little by way of detail on how social learning works. Later models made explicit various forms it could take. The initial model posited a like-meets-like bias in the pairing of players but did not fill in this assumption—it took the form of a correlation parameter. In later models, correlation was generated endogenously due to the fact that interactions occurred within a network and updating was neighborhood based. Thus, the newer models elaborated on the earlier model, proposing mechanisms for how its features arise. Significant details were added, but the basic picture was retained: a population of success-seekers in which individuals have a better-than-random chance of meeting their ilk will tend to stabilize at a norm of equal sharing.

---

[13] McKenzie Alexander, *The Structural Evolution of Morality* (New York: Cambridge, 2007)

How much have we learned from these models about the origins of justice? Below, my goal is to answer this question. Let me emphasize that I am construing it as a substantive, largely scientific issue: I want to ask whether the models explain (or begin to explain) the origins of justice in the familiar, primarily causal sense in which scientific theories explain phenomena. My discussion does not rely on an expectation that the models in question yield testable or fine-grained predictions. To expect that would be unreasonable, as we are dealing with a proposal for an explanatory framework, the subject matter of which are highly complex phenomena. But I will take it that the goal is to understand the origins of actual justice in the actual world. My aim is to assess whether the picture offered by game-theoretic models is on the right track.

III. SYMMETRY

Let us first focus on some key symmetry assumptions. Divide the Cake depicts two agents that have identical preferences, no special needs, and no claims on the good they are dividing. Moreover, the good itself is nondescript: it is a windfall with no history and no intended use; none of the agents involved had any role in producing or procuring it; it does not belong to anyone; and there are no effects to distributing it beyond the immediate payoffs involved. The game describes in a simple form a pervasive kind of interpersonal interaction—bargaining for a resource in scarce supply—but the interaction occurs against an asocial background.

These symmetry assumptions greatly affect the kinds of norms that can be represented. For instance, many actual norms of distribution, including a variety of egalitarian norms, take into account the relative contributions of agents—direct or indirect—to the production of the goods being distributed. Many norms also take into account the effects of distributing a good, how needy the distributees are, as well as the ends to which they will use their share.[14] Such attributes of players are not represented in Divide the Cake, and, at least with respect to some of them (for example, neediness), it is unclear that a model of this sort can represent the relevant facts. Moreover, it is highly doubtful that an evolutionary model that incorporated these aspects would yield an egalitarian result similar to equal splitting.

---

[14] Empirical results in psychology shore up these casual reflections. See Menahem Yaari and Maya Bar-Hillel, "On Dividing Justly," *Social Choice and Welfare*, I, 1 (1981): 1–24; Bar-Hillel and Yaari, "Judgments of Distributive Justice," and Richard J. Harris, "Two Insights Occasioned by Attempts to Pin Down the Equity Formula," both in Barbara A. Mellers and Jonathan Baron, eds., *Psychological Perspectives on Justice* (New York: Cambridge, 1993), pp. 55–86, 32–54.

Thus, while the equal split might be representative of egalitarian rules of justice, its degree of generality is fairly limited. Even if we were to take the Divide the Cake models to provide a full explanation of the tendency of laboratory subjects to demand one-half and the concomitant intuitive pull of the equal-split solution, it is far from clear that we can extrapolate from this to other cases where distributive problems arise, even cases in which an egalitarian solution is salient.

IV. TYPES OF LEARNING

Now let us set aside worries about generality and look at a more fundamental explanatory question. The evolutionary game-theoretic models that are our focus portray moral change as a cultural evolutionary process. This type of populational change is not biological, but it bears important similarities to evolution by natural selection.[15] The units of cultural moral evolution (on the game-theoretic picture) are moral "strategies" which have consequences for population-level ways of dividing goods. The process is driven by the differential adoption of strategies at the individual level, where learning serves in the role of heredity. Thus we have an analogue of "heritable variation in fitness"—to use Richard Lewontin's phrase—the basic pattern of a Darwinian process of change. The key to assessing the explanatory value of the game-theoretic framework lies in determining whether it is plausible that a Darwinian dynamic underlies cultural moral change. To do so, we need to take a closer look at social learning and some different forms it may take.

Success-based Learning. On the picture underlying game-theoretic models, social learning is the engine of moral change. The central idea is that moral learning is driven by success in strategic interactions. The fact that learning results in the spread of successful strategies is what gives the model its Darwinian character. 'Learning' is used here in a generalized fashion to refer to ways in which an individual adopts new behavioral routines. What distinguishes different learning rules are the biases they introduce into the adoption of new behaviors. In success-based learning, as I will call it (or simply

---

[15] Many (arguably most) writers in this area nowadays would reject the idea that the contours of moral norms in present-day societies can be explained purely, or even primarily, as the result of biological evolution.

"success learning"), the adoption of new behaviors is biased towards those that increase payoff. That is, individuals adopt new behaviors in a payoff-driven fashion, seeking strategies that will increase their gains in future rounds of the game. Here payoff is understood as a measure of the subjective valuation of the alternatives, from the point of view of the players—broadly akin to utility. The idea that learning is success based is pivotal to the game-theoretic framework and to the explanation of justice here at issue, so it is worth going into more detail.

The initial model employed the replicator equation. As noted above, this equation of change has strategies spread as a function of how well they are doing relative to the population average. Strategies that do better than average proliferate. The replicator equation can receive a cultural interpretation. The idea is that if individual agents tend to mimic those who perform better, strategies that do better will get mimicked at a higher rate across the population. The replicator equation thus represents the aggregate effects of individual-level success learning. In the later agent-based models, success learning is given a more explicit form. Individuals are represented severally, so for every individual the model must specify an update rule. This allows the injection of some real-life complexity into update rules. Skyrms and Alexander used a characteristic success-based rule—"imitate your best neighbor." Other success-based imitation rules include "imitate proportional to success" and "imitate the winner." But success learning is not specially tied to imitation. Consider "best response": this rule has the player monitor the behavior of all her neighbors and then calculate which strategy will give the highest overall payoff in the next move, assuming neighbors will continue to behave as they have been. "Best response" is a high-rationality mechanism compared with imitating more successful players. Less sophisticated rules also exist. In reinforcement learning, for instance, a player chooses her strategy based on the cumulative payoff it has yielded in the past.[16]

What I have called success learning is therefore a bias in the direction of adopting behaviors associated with specific expectations about success—those that are likely to yield higher payoffs in the future or that have done so for others (the latter, of course, is often a good proxy for the former). The appeal to success-based rules represents a picture in which agents are driven to change their behavior by directly attending to the relation between their success and that of their peers.

---

[16] For a discussion of these and other learning rules, see Young, *Strategic Learning and Its Limits* (New York: Oxford, 2004); Drew Fudenberg and David K. Levine, *The Theory of Learning in Games* (Cambridge: MIT, 1998).

To better understand what is at stake, it will help to consider other types of social learning. Here are two important ones.

Source-based learning involves a bias in favor of particular role players in the learner's environment—parents, teachers, superiors, celebrities. In source-based learning an individual's adoption of behaviors is guided by the fact that other individuals, those who occupy a particular role, exhibit this behavior. In recent work, Kevin Laland has referred to this as "who" learning.[17] There may be a (biological) evolutionary explanation for why children and young adults tend to emulate parents or other salient individuals in their surroundings. The resultant sharing of knowledge or the strengthening of in-group cohesion may have enhanced ancestral fitness. But models of cultural evolution start, as it were, where biological evolution ends.[18] Their concern is with figuring out how the process of cultural change looks *given* a preexisting cognitive toolkit, including patterns of learning and social adaptation. From the perspective of modeling a process of cultural change, source-based learning behaves very differently from success learning.

In source learning one follows in the footsteps of salient individuals such as parents despite or irrespective of their success. It is easy to translate this idea into specific modeling suggestions. Consider a network model of the sort discussed above. In any given neighborhood one individual is designated the leader. A simple (and rather boring) "follow the leader" rule has all individuals adopt the behavior of their local leader. Probabilistic leader-following is a reasonable next step. A plurality of leaders, with different locations and strengths, could be introduced, and so on.

Frequency-based learning involves another type of bias in the adoption of behaviors. Here an individual adopts behaviors as a function of how common they are in her environment. The clearest examples involve herd phenomena, or conformism more generally. Of course, as with source-based learning, there may be evolutionary explanations for frequency biases.[19] But from the point of view of modeling a social dynamic, frequency-based learning differs greatly from success learning.

---

[17] Kevin Laland, "Social Learning Strategies," *Learning and Behavior*, XXXII, 1 (2004): 4–14.

[18] It is likely that both processes occur at once. But it is also likely that their time scales are substantially different. For this reason one can treat the biological as fixed when studying cultural evolution.

[19] See, for example, Jeremy Kendal, Luc-Alain Giraldeau, and Laland, "The Evolution of Social Learning Rules: Payoff-Biased and Frequency Dependent-Biased Transmission," *Journal of Theoretical Biology*, CCLX, 2 (September 2009): 210–19.

As with source learning, the idea of frequency bias can be readily modeled. It is also apparent that processes in which herd effects play an important role will tend to behave differently than success-based processes. The initial configuration of a population will matter more, and one expects to find snowball effects owing to the self-reinforcing nature of conformism and herd phenomena. Frequency bias also leads to populations that are less susceptible to small perturbations, erratic flips in an individual's behavior, trial and error, and the like. When abundance is what determines a strategy's fate, the appearance of a small number of outliers is less likely to make a significant impact.

V. JUSTIFYING SUCCESS LEARNING?

I have presented success learning, source learning, and frequency bias in their pure forms. It is natural to suppose that all three play a role in moral learning, in addition to other factors (for example, random experimentation and inertia).[20] Highlighting the differences among types of learning brings out the fact that in evolutionary game-theoretic modeling an assumption is being made—namely, that moral learning is purely success based (call this the "assumption of success learning"). The explanatory value of the Divide the Cake model and others like it depends on this assumption. How plausible is it?

There are two kinds of ways, I think, of justifying the assumption of success learning which are consistent with the models having substantial explanatory value. First, one could provide empirical evidence that moral learning is largely success based. Second, one could show that it does not matter: the results of the model do not depend very strongly on the assumption. Neither argument has been made, as far as I know, in the context of investigating the origin of morality. If made, I think neither would succeed.

The empirical literature on moral education and development is not couched in terms of the categories I have discussed. But there is ample evidence that social learning in general, and moral learning in particular, involves source and frequency biases. Many psychologists studying moral development take parental guidance to be the key to the acquisition of moral norms.[21] Cognitive-psychological experiments demonstrate that people differentially pick up and retain information from designated sources in their environment.[22] Psychologist Joseph Henrich and anthropologist Francisco Gil-White have emphasized the importance of such mechanisms for cultural change in general.[23] They discuss evidence showing that prominent individuals in a community tend be listened to and emulated, both with respect to factual beliefs and standards of conduct. Interactions with socially prominent individuals have a particular behavioral-cognitive signature, too. Anthropological observations in traditional communities show that elders are observed more closely and that information from and about them is retained better.[24] More generally, it is highly plausible that power asymmetries and hierarchical organization are produced and sustained by their influences on the conduct of individuals. Such hierarchical power structures occur within small-scale communities, within the family, and in the context of specific institutions in politically organized societies. In all these contexts, top-down influences clearly play a role in shaping individual moral conduct—either via mechanisms of emulation and internalization, or through fear and compliance. The extent of these kinds of effects is hard to assess, but the range, strength, and frequency of asymmetries in social standing suggest that such influences are important indeed.[25]

---

[20] How these factors interact and how to model their interaction are interesting questions in their own right. It is possible their effects are additive, in which case each factor can be assigned a weight in an overall sum. More complex possibilities include default/threshold rules, such as "follow the leader, unless you gain more than x by doing otherwise" or "maximize payoff, unless you find yourself in too small a minority."

[21] See essays collected in Jerome Kagan and Sharon Lamb, eds., *The Emergence of Morality in Young Children* (Chicago: University Press, 1983); and Judith G. Smetana and Melanie Killen, eds., *Handbook of Moral Development* (Mahwah, NJ: Laurence Erlbaum, 2006).

The second way one could justify the assumption of success learning is via robustness analysis. Roughly put, robustness analysis shows how well a result holds up in a variety of set-ups, thereby demonstrating which factors make a difference to it.[26] Modeling the pressure and volume of a gas (of fixed volume and temperature) while varying the shape of the chamber is a way of examining whether the geometry of the chamber is a difference-making factor (for the dependence of volume on pressure). If varying the geometry has no effect on the relationship between pressure and volume then this is evidence that geometry is not a relevant difference maker.[27] Analogously, if we substituted for the success-based learning rules used in Divide the Cake models a variety of rules like the ones discussed in the previous section, and if, doing so, we still observed a strong tendency to arrive at an equal split, this would tell us that equal splitting is not heavily dependent on the assumption of success learning. To the best of my knowledge, such analysis has not been done. Were it to be done, I think it rather unlikely that equal splitting would be shown to be robust across different learning rules. If initial conditions include a substantial number of non-equal-splitters, it seems that a conformist bias would land the population in one of the common strategies. If source learning plays a significant role, it will matter greatly what a small number of leader nodes in the social network do, less so the strategies of the mass of followers. Thus I think it is doubtful that one can justify the focus on success learning by arguing that little causal information is lost by the exclusion of other types of learning.

---

[22] Thomas Holtgraves, John Srull, and Janet Socall, "Conversational Memory: The Effects of Speaker Status for the Assertiveness of Conversation Remarks," *Journal of Personality and Social Psychology*, LVI, 2 (February 1989): 149–60.

[23] Joseph Henrich and Francisco Gil-White, "The Evolution of Prestige Freely Conferred as a Mechanism for Enhancing the Benefits of Cultural Transmission," *Evolution and Human Behavior*, XXII, 3 (May 2001): 165–96.

[24] *Ibid*

[25] There are also theoretical reasons for thinking that source-specific influences are biologically adaptive. See Robert Boyd and Peter J. Richerson, *Culture and the Evolutionary Process* (Chicago University Press, 1985); as well as Sabine Coussi-Korbel and Dorothy M. Fragaszy, "On the Relation between Social Dynamics and Social Learning," *Animal Behavior*, L, 6 (1995): 1441–53.

[26] For a philosophical account see Michael Weisberg, "Robustness Analysis," *Philosophy of Science*, LXXIII, 5 (2006): 730–42.

* * *

A different kind of reply to the criticism I have made is that it is premature. Models of the evolution of morality, and the models we have looked at in particular, are at an early stage. Like most modeling, its beginnings are simple and limited. The models target a restricted set of phenomena, make simplifying assumptions, and aim for a partial fit with the world. Specifically, one might concede that the assumption of success learning represents an oversimplification but argue that it is merely a natural starting point, not an end to the investigation. Criticisms of it are founded on an unreasonable expectation that the model take the full complexity of real-life social phenomena into account. Such "start simple" responses are encountered fairly often in discussions of models. Impatient outsiders seek completeness while modelers exhibit a more pragmatic, one-factor-at-a-time approach, seeking a good understanding of simple scenarios before moving to more complex ones.

The "start simple" response is important and, as far as it goes, correct. But it cuts both ways: if models of the evolution of morality are at an early stage in which it is hard, as yet, to justify the idealizing assumptions being made, then optimism about the explanatory value of this work should be kept in check. Indeed, caution should be taken with respect to the basic message of work on the evolution of morality—namely, that moral conduct evolves, in a way closely analogous to how biological traits evolve. The assumption of success learning is what gives the models their Darwinian character; it is what ensures that proliferation tracks payoff. Doubts concerning this assumption undermine the basic picture of morality as evolving through a form of Darwinian cultural selection.

---

[27] Of course, one needs to ascertain that geometry was varied in the right ways and that other significant factors have been held fixed.

Success learning appears as a natural starting point if and to the extent that one thinks it is likely that morality evolves by something like natural selection. It is sometimes thought that if there is to be a scientific account of cultural change and the stabilization of norms, then this account must in some important sense be akin to Darwinian natural selection. But it is important to see that is not so. What is characteristic of Darwinian dynamics is that the differential spread of types in a population tracks success. In biology, success equals increased offspring; in cultural evolution, it translates into the rate at which individuals adopt a behavior or an idea. Some populational trends are not Darwinian in character. Bank runs, dress fads, and other herd phenomena are not primarily Darwinian. Nor are top-down changes, such as the near cessation of religious activities in the former U.S.S.R. Herd behavior is not driven by success but by a sheer snowball effect: people emulate what appears to be most common. Top-down changes are sometimes driven by success, but only that of the individuals at the top. Thus, neither process is Darwinian.[28] Now, importantly, moral change might be more like herd behavior or top-down changes than it is like natural selection. It would be so if frequency effects and source learning are rife. If this is the case, the idea that morality evolves is fundamentally incorrect. In this sense the most basic message of work on the evolution of morality will not have been borne out.

VI. MODES OF PROGRESS

The discussion of simple beginnings directs our attention to the diachronic aspect of model-building. In modeling one starts simple but, ordinarily, seeks to add complexity as time goes by. It is to the diachronic aspect of modeling the evolution of morality, and of modeling more generally, that I want to turn in this final section.

I have talked throughout in terms of models and modeling, but I have not yet said much about model-building as such. As I understand it, modeling is a part of scientific (and perhaps philosophical) theorizing characterized by a pragmatic approach. Instead of proposing a broad, unified theory of a certain domain of phenomena, modelers proceed piecemeal, offering a family of partial theoretical proposals in hopes of breaking down a large and complex problem into tractable bits. Theorizing about the evolution of morality is in this respect a typical case of modeling. In general, models are usually local in scope and tend to contain significant idealization—assumptions that are known to be false are introduced in order to simplify the analysis and isolate causal factors.

---

[28] For a related discussion, see Peter Godfrey-Smith, *Darwinian Populations and Natural Selection* (New York: Oxford, 2009), especially chapter 8, pp. 147–64.

One important kind of model-based inquiry involves the indirect investigation of a real-world phenomenon ("the target") via the analysis of a hypothetical construct ("the model").[29] Substantial parts of theory in evolutionary biology and in economics have this character. If model and target sufficiently resemble one another in their causal structure and/or in their input-output behavior, the analysis of the model can provide information about the target. Indirectness is often indispensable, in that it affords a separation of the analysis of theoretical possibilities from an evaluation of their empirical plausibility. In the sciences of complex phenomena, theorizing that is overly focused on the real world can hamper understanding rather than promote it. But indirectness can lead theorists astray precisely because it legitimizes a looser fit between theory and reality. One goal of the present paper has been to evaluate the explanatory merits of game-theoretic models of the evolution of morality. Another is to connect this evaluation to considerations that apply more generally to the process of modeling and its pitfalls.

Let me distinguish two modes in which progress in indirect modeling can be made. Both have epistemic significance, but of a different sort. Both characteristically involve moving from simple beginnings to more complex analyses. One mode of progress can be called "target oriented." Here the move towards complexity is sensitive to factors that affect the behavior of the real-world target, elements that make a causal difference to it. In the simplest (and perhaps ideal) case, a model starts by examining one or a small number of important factors, typically in a simplified way, and initial results are obtained. As time goes by an effort is made to incorporate further factors and to gauge the importance of the factors included in the initial proposal. Additional elements are either brought in or discovered to be irrelevant. If this kind of effort goes well then the result is greater understanding of a real-world target: progressively better knowledge about the causal process underlying the explanandum. We can contrast target-oriented progress with an "internal" mode of progress, which is more conceptual in spirit. Here a particular picture or set of assumptions is laid out and explored in its own terms. Detail is added primarily to refine the understanding of elements that are present to begin with. Internal progress is made to the extent that the initial construct and its behavior are well understood, and that refinements are seen to either bear out and illuminate the initial results or complicate them in interesting ways. Thus, in a target-oriented mode a modeling endeavor makes progress by incrementally adding causal information. This can be done via experimental testing and empirical observation or via robustness analysis, but the crucial point is that the work is regulated primarily by its empirical adequacy. In contrast, in an internal mode one explores the subtleties of a constructed set-up, but this is done largely independently of the actual causal relevance of the set-up being explored.

---

[29] I draw here, with some modifications, on a framework developed by Ronald Giere in *Explaining Science: A Cognitive Approach* (Chicago: University Press, 1988), and more recently by Godfrey-Smith, "The Strategy of Model-Based Science," *Biology and Philosophy*, XXI, 5 (November 2006): 725–40; and Weisberg, "Who Is a Modeler?" *British Journal for the Philosophy of Science*, LVIII, 2 (2007): 207–33.

Skyrms's initial model depicted a cultural process akin to evolution by natural selection and showed that it could lead to an egalitarian rule of justice. The underlying causal picture was of individuals haggling over a good and adapting their behavior to achieve greater bargaining success. Skyrms found that correlated pairing of players was vital. Later work was primarily devoted to filling in the initial picture. Network structure, which generated the correlation among players endogenously, was introduced, as were specific success-based learning rules, which gave concrete form to the statistical trend embodied in the replicator equation. This amounted to real progress, I think, but progress of the internal kind; it was an exploration of the same sort of set-up that was posited at the start. It was not progress of the target-oriented, causally informative variety. Little or no attempt was made to provide empirical grounding for the idea that "straight" bargaining is the right way to capture justice or for the key simplifying assumption that moral learning is success based. Nor was there a theoretical attempt to examine how much of an effect these assumptions have on the basic message that morality changes through a cultural process akin to evolution by natural selection.

Now, some statements made by modelers in this area suggest otherwise. They assert that substantial explanatory progress has been made. Skyrms and William Harms, for instance, state that in the past "the nature and source of [moral] standards have remained something of a mystery," but "[r]ecent work on the evolution of norms has changed this picture dramatically."[30] In light of the preceding discussion, I think that what is going on here, at least some of the time, is that internal progress is being mistaken for target-oriented progress. This is a mistake that, I want to suggest, is connected with theorizing in the indirect style and the two modes of progress I have outlined. For while they differ methodologically and epistemically, internal progress and target-oriented progress often *look and feel* quite similar from the point of view of model development. Both characteristically involve incremental concretization—adding detail, positing specific mechanisms, substituting aggregates by explicit representation of individuals, and so on. This phenomenological similarity can be mistaken for epistemic parity, with the result that research that has taken a largely conceptual route appears to have yielded progress in answering explanatory questions about the empirical world.

---

[30] William Harms and Skyrms, "Evolution of Moral Norms," in Michael Ruse, ed., *The Oxford Handbook of Philosophy of Biology* (New York: Oxford, 2008), pp. 434–50, at p. 434.

While my argument only licenses conclusions about models of the evolution of morality, it is important to recognize that this *kind* of situation can, in principle, occur in other instances of indirect modeling. That is to say, there is a type of error that can arise when an indirect modeling project, with its associated phenomenology, is developed in one way but interpreted in another. Notice that this diagnosis differs from a suggestion that is commonly made when models are thought to float free from reality, namely, that mathematical elegance or computational tractability are overvalued relative to empirical adequacy. Such complaints have been made recently in extra-academic settings against macroeconomists and string theorists.[31] It is possible that tractability and elegance occupy too prominent a role in these instances. But it often seems that accusations of this sort present modelers in an unreasonably crude light. On the present proposal, the error is subtler than that. It does not involve a simple conflation of mathematical appeal and truth. Rather, it arises when one fails to distinguish between two kinds of motion towards greater complexity which are genuinely similar in some respects but differ in their epistemic import. Whether and to what extent this has occurred in any particular instance depends on the details. A healthy skepticism would counsel in favor of bearing the possibility in mind.

ARNON LEVY

The Van Leer Jerusalem Institute

---

[31] With respect to the former see, for example, Paul Krugman, "How Did Economists Get It So Wrong?" *The New York Times Magazine*, September 2, 2009. (I am indebted to Peter Godfrey-Smith for drawing my attention to this aspect of Krugman's article.) For the latter, see Lee Smolin, *The Trouble with Physics* (New York: Haughton Mifflin, 2006).

A MOSQUITO BITE AGAINST THE ENACTIVE APPROACH TO BODILY EXPERIENCES*

The enactive approach to bodily experiences rejects the classical computational dichotomy between perception and action.[1] Perception is not merely a means to action and action a means to perception. Instead, proponents of the enactive approach argue for a stronger relation of constitutive interdependence. On this view, perceptual content constitutively depends on law-like relationships that hold between sensory input and motor output. Hence, perceptual experiences are said to be inseparable from the perceiver's bodily activities.

Enactivists take touch as paradigmatic of perception, because the act of exploring the shape of an object with one's fingers is said to constitutively determine the content of one's tactile perception.

> All perception is touch-like in this way: perceptual experience acquires content thanks to our possession of bodily skills. What we perceive is determined by what we do (or what we know how to do).[2]

> Touch acquires spatial content—comes to represent spatial qualities—thanks to the ways touch is linked to movement and to our implicit understanding of the relevant tactile-motor dependencies governing our interaction with objects.[3]

Although vision has received the most attention, it seems that the enactive approach is more likely to be true of tactile experiences than those of any other sense (such as gustatory experiences). For instance, in haptic touch, the tactile system does not passively receive bodily information, but consists in exploratory activities. However, what may be true for haptic touch, which is intrinsically active, may be more controversial for passive touch (for example, the sensation of a mosquito on your skin). If touch is to be the model for all perception, or

---

* I am grateful to Ned Block and Pierre Jacob for their useful comments and their support.

[1] This view has been defended by, among others, Susan Hurley in *Consciousness in Action* (Cambridge: Harvard, 1998), Kevin O'Regan and Alva Noë, "A Sensorimotor Account of Vision and Visual Consciousness," *Behavioral and Brain Sciences*, xxiv, 5 (October 2001): 939–1031; Noë, *Action in Perception* (Cambridge: MIT, 2004); Evan Thompson, "Sensorimotor Subjectivity and the Enactive Approach to Experience," *Phenomenology and the Cognitive Sciences*, iv, 4 (2005): 407–27

[2] Noë, *Action in Perception*, p. 1

[3] *Ibid.*, p. 205.

0022-362X/11/0804/188–204

at least the less controversial application of the enactive approach, then it calls for a more thorough examination.

After drawing the agenda for an enactive view of tactile experiences, I shall highlight the difficulties it has to face, both conceptual and empirical. I shall conclude that what we do does not determine what we feel.

I. THE AGENDA FOR AN ENACTIVE APPROACH TO TACTILE EXPERIENCES

At the core of the enactive approach is the dynamic, constitutive relationship between sensory input and motor output. On this view, we are able to track their interdependence. This ability is grounded in the procedural knowledge of how the way we move affects the sensory signals we receive (or how the movement of objects will affect those signals). In this sense, Hurley talks of "active perception," and Noë describes perception as an "embodied activity."

One should, however, acknowledge the diversity of views that enactivism includes. Indeed, there is little agreement among the proponents of the enactive approach beyond the claim that perception and action are interdependent. One noteworthy difference is the variety of *explananda* among the enactive views. There are at least three phenomena that are thought to require an account in enactive terms: (i) the individuation of perceptual modalities (for example, sight versus audition); (ii) perceptual phenomenology (that is, what it is like to perceive); and (iii) perceptual content (that is, representational content). First, it has been argued that each sensory modality depends on its own, specific patterns of dynamic interdependence between sensory stimulation and movements. The set of patterns differs depending on whether one is seeing or touching an object, for instance, and as such would individuate the various sensory modalities, here, distinguishing sight from audition. Second, some versions of the enactive approach claim to account for consciousness, and more precisely for the qualitative phenomenology of perceptual experiences. For instance, what it is like to see a red tomato would consist in the procedural knowledge of the sensory effects of one's movements relative to the light. Finally, most enactive theories aim to explain the intentional content of perceptual experiences, and particularly how perceptual experiences represent spatial properties. Seeing that the table is round is to know that it looks more or less elliptical depending on where one is in relation to it.

Here I am not interested in the idiosyncrasies of the various enactive approaches.[4] Rather, I shall propose different ways to spell out the enactive view of tactile experiences and focus on the most plausible ones. Following the distinctions above, an enactivist for tactile experiences can pursue three aims: (i) to individuate the tactile modality (for example, in contrast to a visual one); (ii) to explain the phenomenology of tactile sensations (for example, the conscious sensation of pressure on my skin); (iii) to explain the spatial content of tactile experiences (for example, the sensation that there is a round object touching my palm). In the last case, the enactivist will have to take into account the duality of the content of tactile experiences. In contrast to other sensory modalities, touch is both about the external world with which one is in contact and about one's body in contact with the external world. Hence, an enactive account of tactile experiences must accommodate both exteroceptive content and interoceptive content.

---

[4] For a review of the differences between versions of the enactive theory of vision, see Nivedita Gangopadhyay and Julian Kiverstein, "Enactivism and the Unity of Perception and Action," *Topoi*, XXVIII, 1 (2009): 63–73.

To add to the complexity, touch is a very comprehensive term, so much so that one may even deny that it refers to a unique sensory modality. On the one hand, there is active touch. It primarily is defined by movements of the body part that is in contact with an object. The movements can be of three types. Some are passive movements (for example, your hand is moved by another individual). Some are exploratory movements, voluntarily initiated by the perceiver (for example, you follow the contours of an object). Some are effortful movements (for example, you lift and manipulate an object with muscular effort). On the other hand, there is passive touch, which is merely cutaneous. Here again, one can distinguish between static passive touch (for example, the pressure of the cat on your knees) and kinematic passive touch (for example, the motion of the mosquito on your skin).

In virtue of their interoceptive content, tactile experiences are partly bodily experiences, or in more enactive terms, *embodied*. However, what matters for the enactive approach is not so much the embodiment of perception as its intrinsic relation to action. It is controversial whether tactile experiences consist in an embodied *activity*. True, in haptic touch, I move my body to retrieve information about an object when needed, and the movements of my hand over a surface induce shifting sensations. But can one generalize the enactive account of haptic experiences to any kind of tactile experiences, and even more generally, to bodily experiences? How can so-called passive touch be *active*?

Here I shall attempt to apply the enactive approach to instantaneous passive touch, which seems to be the most challenging case. Let us imagine that I consciously feel a short, nonpainful tap on my knee. The content of my experience is not very rich. I may not even be able to know what touched me. Yet, I feel that I was touched briefly *on my knee*, and not anywhere else. What grounds my phenomenology and the spatial content of my tactile experience? If one succeeds in offering an enactive account of such an impoverished tactile experience, then there is a good chance that one can offer a more general enactive account of bodily experiences. Let us now see how this might work. We shall see that the task is not easy.

II. TACTILE-MOTOR EXPECTATIONS

Visual experiences are said to depend on eye movements, and haptic touch on exploratory hand movements. But I do not need to move to have a tactile experience, as in the case in passive touch. Yet, a proponent of the enactive approach may claim that even in passive touch, I master the procedural knowledge of the tactile consequences of potential movements (that is, sensorimotor contingencies). What matters is not the bodily exercise of procedural knowledge, but its mere possession. For instance, itching sensations would consist in the knowledge that if one scratches where one has been bitten by a mosquito, the stimulation will decrease. One does not need to scratch one's leg; it suffices that one knows the consequences of doing so. The same could be said for tactile experiences. What would constitute tactile experiences is not what you do, but what you expect to feel if you do it. But what kind of procedural knowledge is constitutive of passive touch?

Let me start with two possible tactile-motor expectations.[5]

(1) If I remove my knee from the object with which it is in contact, I shall stop receiving tactile signals.
(2) If I remove the object with my hand, I shall stop receiving tactile signals.

As basic and uncontroversial as these tactile-motor expectations are, they raise several worries. The closest equivalent in vision would be the expectation that if I close my eyes, I will no longer receive visual signals. Nobody in the computational community would deny this simple relation of causal dependence between vision and eye opening/closing, or between touch and bodily contact. However, not all causal relations between touch and action are of interest for the enactive view. According to Hurley, the interdependence of perception and action can be of two kinds. It is instrumental when action is just a means to perception, and it is noninstrumental when perception wholly or partly consists in action. Only the latter kind is relevant for the enactive approach. The question, then, is which kind of sensorimotor expectations (1) and (2) are: instrumental or noninstrumental? In vision, the movement of eye opening/closing makes the visual experience possible or impossible, but this is a feature of any kind of visual experience. I could see a red tomato or a yellow lemon for all it matters. As such, this sensorimotor expectation cannot determine the specific phenomenology or the specific content of the visual experience that results from opening my eyes. Similarly, in touch, the movement of separating the body part from the object stops the tactile experience, but it would stop any kind of tactile experience, whether I feel a rough texture or a smooth one. Hence, it cannot be an individuative condition of the phenomenology of my tactile experience, nor of its descriptive content. At most, this type of sensorimotor contingency may individuate the sensory modality: vision would constitutively require stopping the flow of information by closing the eyes, and touch would constitutively require stopping the flow of information by withdrawing the body part in contact with the object. Yet, it seems that most enactive proponents want more.

---

[5] Many questions have been raised about the specific nature of sensorimotor knowledge. Here I shall not insist on these general difficulties about the enactive account, but focus on specific problems raised by tactile-motor knowledge. For a full review in the case of visual experiences, see the commentaries in O'Regan and Noë, *op. cit.*

They may argue, for instance, that the tactile-motor contingency (1) can individuate at least the bodily spatial content, which is at the core of our interests. Indeed, it is only if I remove my knee from the object with which it is in contact that I can expect the tactile signal to stop, and not any other body part. The tactile-motor contingency is body-part specific, and as such might provide the bodily spatial content. But how do I know that it is my knee I need to move? There is a threat of circularity similar to the one Jacob highlighted in the case of vision.[6] The enactive account claims that visuo-motor knowledge consists in the knowledge of how one's movements toward a visually seen object affect the way one perceives the object. But Jacob notices that in order to act upon an object, one first needs to be able to single it out, to pick it out from other objects. Hence, visuo-motor knowledge cannot allow one to individuate the relevant object. There must be another part to the story, which cannot be told in enactivist terms. Similarly, in the tactile case, it is not clear how one can individuate the body part on the basis of tactile-motor contingencies without falling into circularity. Tactile-motor knowledge requires a prior and independent way of singling out the body part that is in contact with the object, if one wants to avoid a circular account of bodily spatial content.

Tactile-motor expectations (1) and (2) do not work if they are supposed to constitute the phenomenology or the content of tactile experiences. But the problem is more fundamental, and not limited to these two examples. To give an enactive account of the content of tactile experiences, especially the interoceptive content, is indeed difficult. On some enactive views, what constitutes the content of perceptual experience is the procedural knowledge that gives access to objective properties based on the relation between the experience of perspectival properties and bodily activities. For instance, the table is round (objective property of the stimulus independent of the perceiver), but it appears elliptical to you (perspectival property, the way the stimulus appears to the perceiver). According to this enactive theory, your visual experience represents the table as round because you know how its perspectival property will change depending on how you move relative to the object (for example, you get closer to the table). However, on this account it is hard to see how there could be any kind of dependency between touch and action, as far as the interoceptive content of touch is concerned. When I feel the tap on my knee, it does not seem to make sense to distinguish between a perspectival bodily property and an objective bodily property, if perspectival properties depend on the spatial relation between the perceiver and the object. If the object is one's own body, then it cannot be blocked from tactile perception; nor can one change one's spatial relation to it. In other words, one cannot change the spatial conditions under which one's own body is felt. And if one cannot change the spatial conditions, then there is no corresponding tactile-motor knowledge of the way the perspectival properties will change depending on the way one moves relative to one's body. Faced with this difficulty, proponents of the enactive view have two choices: either restrict their account to the exteroceptive content of tactile experiences or offer an alternative enactive account of the interoceptive content that differs from the one provided for the exteroceptive content.

---

[6] Pierre Jacob, "The Scope and Limits of the Enactive Approaches to Visual Experiences," *Perception*, xxxvii, 3 (February 2008): 446–61.

Finally, even if the tactile-motor contingencies outlined above were satisfactory, they could not work for every kind of touch, and so they fail to satisfy the enactive agenda. For example, they do not apply easily to *instantaneous* touch. When I receive a brief tap on my knee, I cannot remove my knee from the object (or remove the object from my knee) because the object is already gone. And if I try to do so, there will be no difference in tactile stimulation because the tactile signal has already stopped. More generally, it is difficult to imagine the interdependence of perception and action when perception is so temporary that it does not afford time to act. Most sensorimotor expectations require enduring perceptual experiences to make sense. If the object is no longer present, there can be no sensory consequences. Hence, there can be no sensorimotor contingencies. If this is true, then there is no possible enactive account of instantaneous tactile experiences (or of any other kind of instantaneous perceptual experience).

The only way out is to include not only sensorimotor knowledge, but also mere motor knowledge in the enactive account. One could argue that it does not matter that I am too late in removing my knee from the object that briefly touched me. The action does not even need to be performed; it merely needs to be triggered by the tactile signal. The urge to move would constitute the perceptual experience independently of its bodily execution, its sensory feedback, and the remaining presence of the sensory input that elicited it. Procedural knowledge does not require that the object to which it applies still exists. Let us imagine that you tell me about a restaurant that you liked, which closed recently. If I ask you where the restaurant *was*, you may still know how to get there (although the restaurant is no longer there). Similarly, I can still act toward the location of the tactile stimulus, although the stimulus is no longer available. Feeling that I was briefly touched *on the knee* is to know how to get to where I was touched.[7] On this view, the spatial content of the tactile experience is determined by the procedural knowledge of how to get to the bodily location of the tactile stimulation. Hereafter, I shall refer to this type of procedural knowledge as "spatial know-how."[8] An account in terms of spatial know-how sounds plausible as a working hypothesis for the enactive approach for two reasons.[9] On the one hand, this kind of procedural knowledge seems to be consistent with the enactive view.[10]

---

[7] If I reach the body location that was touched and put my finger on it, then I shall receive tactile signals from the same place. In this latter case, there are tactile consequences

[8] For a detailed account of the notion of bodily know-how, see Charles Siewert, "Attention and Sensorimotor Intentionality," in David Woodruff Smith and Amie L. Thomasson, eds., *Phenomenology and Philosophy of Mind* (New York: Oxford, 2005), pp. 270–94.

[9] However, one may argue that the spatial content of bodily experiences must be determined by the movements performed by the specific body part that is experienced. For instance, the tactile experience of my knee being touched should be determined by the movements of my knee only. On the contrary, pointing and reaching movements are performed with a body part that is not represented in the tactile content, namely, the hand. Should this type of movement be discounted for this reason? This would be too restrictive. Most of the actions that one can perform towards one's body are hand movements, and if one discounts them as irrelevant, then there are not many movements left over. Furthermore, there are some body parts that are not movable, and yet we can experience them being touched. Therefore, one might suggest that the spatial content of the tactile experience depends on the procedural knowledge of the body part that is touched with respect to the rest of the body.

[10] "What we perceive is determined by...what we know how to do." Noë, *Action in Perception*, p. 1.

On the other hand, I do not see any alternative type of procedural knowledge that could constitute instantaneous passive touch.

To sum up, the link between touch and bodily activity is supposed to be so obvious that it does not need a proper investigation; indeed, it is supposed to help us better understand the (less obvious) enactive view of vision. However, this might be true only for haptic touch, and only for its exteroceptive content. Enactivists still need to give a specific account of the relationship between tactile experiences and action in the case of instantaneous passive touch, and this is not an easy task. They may argue that it is precisely because the relation with action is so weak that the content of such tactile experiences is so poor. That might be true. Nonetheless, for the enactive account to be true, there needs to be at least one type of law-like relation with action if there is to be tactile experience at all. I suggest that the most plausible way to spell out an enactive account of tactile content is in terms of reaching movements made toward the bodily location that was touched. However, we shall see now that this view faces severe difficulties when confronted with the empirical literature on bodily experiences. I shall show that action and tactile experiences are independent, rather than interdependent, and that they can be dissociated in both pathological and normal cases.

III. DISSOCIATING BODILY EXPERIENCES AND SPATIAL KNOW-HOW

One of the most powerful empirical arguments made against the enactive approach to vision comes from the neuropsychological dissociation between optic ataxia and visual agnosia.[11] For example, the visual agnostic patient DF was presented with a set of various squares and rectangles. She was at chance when required to match the width of such simple geometrical forms by scaling the distance between her thumb and index finger (that is, by manual, nonverbal report).[12] By contrast, she was able to grasp the blocks accurately. The reverse pattern of performance was found in optic ataxic patients.

---

[11] For a comprehensive review of the empirical literature that supports the so-called Perception-Action model of vision, see David Milner and Mel Goodale, *The Visual Brain in Action* (New York: Oxford, 1995), and Jacob and Marc Jeannerod, *Ways of Seeing: The Scope and Limits of Visual Cognition* (New York: Oxford, 2003). For an extensive discussion of the relevance of this type of empirical evidence for the enactive approach, see Jacob, *op. cit.*; Ned Block, "Review of Alva Noë, *Action in Perception*," this JOURNAL, CII, 5 (May 2005): 259–72; Andy Clark, *Supersizing the Mind* (New York: Oxford, 2008).

[12] Some hand movements can count as perceptual reports. For instance, it was shown that the underlying processes of scaling the distance between thumb and index finger are very different, depending on whether it is for reporting a perceptual judgment or for the visuomotor task of grasping. Jeannerod, *The Cognitive Neuroscience of Action* (Cambridge: Blackwell, 1997).

These cases show that one can have visual experiences of an object without being able to act towards it, and vice versa. That is, visual awareness can operate independently of visually guided action.

A similar line of evidence against the enactive approach to vision comes from healthy individuals. It has been found that action can be immune to visual illusions such as the Müller-Lyer, Ponzo, Titchener, and Hollow Face illusions. For example, in the Hollow Face Illusion, a concave (or hollow) mask of a face is perceived as a normal, convex (protruding) face. It was found that if asked to quickly flick a magnet off the nose (as if it were a small insect), participants did not direct their finger movements to where the nose appeared to them (that is, to the protruding nose). Instead, they directed their movements to the actual location of the target in the hollow face, which was eight inches away from the apparent location. In other words, the content of the visual experience of the face did not correspond to the visually guided movements directed toward the face. This was taken as evidence that the content of visual experiences (for example, the convex face) is at variance with the content of action-oriented vision (for example, the hollow face). What we see is not determined by what we do, nor what we know how to do.

Is there equivalent empirical evidence against the enactive view of *bodily experiences*? The answer is yes.[13] On the basis of dissociations between tactile and proprioceptive experiences on the one hand and spatial know-how on the other, both in patients and in healthy individuals, I shall argue against the enactive view of (i) the phenomenology of bodily experiences and (ii) the spatial content of bodily experiences.

III.1. Action without Sensation. "But, I don't understand that. You put something there; I do not feel anything and yet I got there with my finger. How does that happen?"[14]

The patient, who was blindfolded, is amazed at her ability to point to where she was touched on her hand, although she felt absolutely no sensation of having been touched. She suffers from what is called "numbsense" in the neuropsychological literature (also "blind touch"). Following cortical or subcortical lesions, patients with numbsense become completely anaesthetized on their right side. They completely lack tactile experience. They are not able to detect, localize, or describe tactile stimuli on their right arm; nor are they able to indicate on a picture of an arm where the stimulus was applied, even in a verbal forced-choice condition. What is interesting is that despite their apparent numbness, they are able to accurately guide their opposite hand to the specific site where they were touched when so instructed, and to their own surprise. Their bodily movements do not improve their performance in verbal localization. When patients are asked to simultaneously verbally localize and point to where they have been touched, both responses become incorrect. In addition, they may be unaware of their arm's position and yet be able to reach it accurately. Interestingly, this type of dissociation was known to Merleau-Ponty. He described the case of the patient Schneider, who was unable to localize where a mosquito was stinging him and yet was able to scratch his leg where he was stung.[15]

---

[13] For numbsense patients, see Jacques Paillard, Francois Michel, and George Stelmach, "Localization without Content: A Tactile Analogue of 'Blind Sight'," *Archives of Neurology*, XL, 9 (1983): 548–51; as well as Yves Rossetti, Gilles Rode, and Dominique Boisson, "Implicit Processing of Somaesthetic Information: A Dissociation between Where and How?" *Neuroreport*, VI, 3 (February 1995): 506–10. For a description of KE and JO, see Helen A. Anema et al., "A Double Dissociation between Somatosensory Processing for Perception and Action," *Neuropsychologia*, XLVII, 6 (May 2009): 1615–20.

[14] Report of a patient suffering from numbsense in Paillard, Michel, and Stelmach, *op. cit.*, p. 550.

When numbsense patients scratch their leg where they were stung, they must have mastered sensorimotor information about reaching their leg, scratching it, and the effects of these actions. Yet, despite this sensorimotor knowledge, they feel nothing. Here we have a clear case of a complete deficit of tactile and proprioceptive phenomenology with preserved spatial know-how. The case of numbsense therefore shows that spatial know-how (and its exercise) *per se* is not a sufficient condition of tactile and proprioceptive phenomenology.

III.2. Action without Perception, Perception without Action. As said earlier, some versions of enactivism aim to explain how perceptual experiences come to represent spatial properties, rather than how they come to have a conscious phenomenology. Hence, one might argue that the evidence provided by the numbsense case is irrelevant for these versions of the enactive view. More challenging for them is the case of the patients KE and JO, who have tactile experiences, but whose tactile experiences are dissociated from spatial know-how in the same way that visual experiences are dissociated in optic ataxia and visual agnosia.

Both KE and JO can consciously feel a touch. In addition, they can accurately report and point to the location of a visual target. Yet, when asked to point to where they were touched, they display a surprising double dissociation. They were asked either to point to where they were touched on the hand (tactile-motor task) or to point to the location of the tactile stimulus on a pictorial map of their hand (abstract and detached perceptual report). It was found that JO failed to point accurately to the hand map, whereas she did point accurately to her own hand. Conversely, KE failed to point accurately to his own hand, but he did point accurately to the hand map.

---

[15] Maurice Merleau-Ponty, *Phénoménologie de la perception* (Paris: Gallimard, 1945).

In neither case does the spatial content of the tactile experiences match the possession (or deficit) of spatial know-how. JO experiences that she is touched, but she cannot report where. She merely experiences the touch "somewhere," with no determinate spatial content. Yet, like numbsense patients, she accurately gets to the location of the touch on her own body. Hence, her spatial know-how does not suffice for providing her tactile experience with spatial content. By contrast, KE experiences and reports that he is touched on the hand at a specific location, but he is unable to get to the location of the touch on his own hand. Hence, there is no spatial know-how that could provide the spatial content of his tactile experience, and the tactile spatial content must have a different ground.[16] The case of JO shows that spatial know-how *per se* is not a sufficient condition of tactile spatial content, whereas the case of KE shows that it is not a necessary condition.

To conclude, these dissociations show that the spatial information in the sensorimotor system can differ from the spatial information in the perceptual system. This leads us to infer that the sensorimotor system and the perceptual system cannot be reduced to one another. They are two distinct systems, one involved in action and the other involved in perceptual experiences. The dissociation between the two types of spatial information specific to each system can be made even more salient in the context of bodily illusions in healthy individuals.

III.3. Illusory Spatial Content. If spatial know-how provides the spatial content of perceptual experiences, then it should explain why this content can become illusory in healthy subjects. However, as said before, it was found that action can be immune to visual illusions. We shall see now that bodily illusions also cannot be explained by mistaken information encoded in spatial know-how. I shall describe here only the Rubber Hand Illusion, but similar results have been found with vibrotactile illusions (vibration of the bicep tendon induces an illusory kinesthetic experience that one's arm is stretching).[17]

---

[16] An enactivist may reply that there is no evidence of a deficit in spatial know-how in KE. There is only evidence of a deficit in spatial motor *performance*, and this does not systematically reveal a deficit in *competence*. If so, tactile spatial content could be accounted for in terms of the supposedly preserved spatial know-how. However, this explanation does not sound plausible. The patient KE has no motor deficit, as shown by his preserved ability to point to a hand map or to neutral visual targets His results can neither be explained by a proprioceptive nor a memory deficit. In the study, the patients' hands were moved passively to a different location, and they were asked after two seconds to return their arms to the previously held position. KE's performance in this proprioceptive task was similar to JO's performance. It thus seems plausible to assume a deficit of spatial know-how in KE, which explains his poor motor performance when pointing to his own hand.

In the Rubber Hand Illusion (RHI), participants sit with their left arm resting on a table, hidden behind a screen. They are asked to fixate on a rubber hand in front of them, and the experimenter simultaneously strokes the participant's hand and the fake hand with two paintbrushes. After a short while, the majority of participants report that they feel the touch of the paintbrush at the location where they see the rubber hand being touched. Even more surprisingly, some participants feel as if the rubber hand were their own hand. At the behavioral level, participants report their hand as closer to the rubber hand than it really is. This is true only when the two hands are in congruent positions and stimulated synchronously.

Now, the enactive view might account for the illusion in terms of erroneous spatial know-how. The explanation might run as follows: participants feel their hand closer to the rubber hand because they were misleadingly induced to expect that if they reach the location close to the rubber hand, they will touch their own hand. However, a recent experiment refutes this explanation of the RHI. It was found that participants accurately directed their opposite hand to the real location of their own hand that was touched, and not to the illusory location that they reported. Their reaching movements were not sensitive to the spatial illusion. Similarly, when they performed the reverse movement (that is, when they directed their touched hand toward their opposite hand), they had the right know-how of where their hand was. The complete absence of illusion was also confirmed in a bimanual task, where participants had to grasp a stick in front of them. Again, their bodily movements revealed accurate spatial information about the relationship between their two hands as well as the use of spatial know-how of their correct locations. However, when participants were asked a second time to make a perceptual judgment

---

[17] For a description of the standard Rubber Hand Illusion, see Matthew Botvinick and Jonathan Cohen, "Rubber Hands 'Feel' Touch that Eyes See," *Nature*, CCCXCI (February 19, 1998). 756 For a description of the standard vibrotactile illusion, see James R. Lackner and Paul A. DiZio, "Aspects of Body Self-Calibration," *Trends in Cognitive Sciences*, IV, 7 (July 1, 2000): 279–88. The immunity of action to the vibrotactile illusion was shown by M. P. Kammers, I. J. van der Ham, and H. C. Dijkerman, "Dissociating Body Representations in Healthy Individuals: Differential Effects of a Kinaesthetic Illusion on Perception and Action," *Neuropsychologia*, XLIV, 12 (2006): 2430–36. The immunity of action to the Rubber Hand Illusion was shown by Kammers, F. de Vignemont, L. Verhagen, and Dijkerman, "The Rubber Hand Illusion in Action," *Neuropsychologia*, XLVII, 1 (January 2009): 204–1[illegible]

CENTRAL LIBRARY UNIVERSITY OF CALCUTTA ALIPORE CAMPUS

about the location of their touched hand after having moved, they were still sensitive to the RHI, and they still experienced their hand as closer to the rubber hand than it was.

Although participants experienced their hand as closer to the rubber hand than it was, they knew how to reach it, and they knew where it was when asked to move it. Hence, the RHI cannot be explained by inaccurate spatial know-how. The illusory spatial content of the bodily experience is not determined by the illusion-immune spatial know-how, which retains an accurate sense of the position of the touched hand.

To sum up, in healthy individuals, like in patients, the spatial content of tactile and proprioceptive experiences can be dissociated from the spatial information encoded in spatial know-how used to guide reaching and pointing movements, such that one can be accurate in one and not the other. Numbsense patients even illustrate the possibility of preserved know-how with no associated tactile sensation. Hence, it seems that spatial know-how, as recruited by reaching, pointing, and grasping movements, is neither necessary nor sufficient for bodily experiences.

IV. THE DILEMMA

In touch, like in vision, these various cases raise interesting questions for the enactive approach. It seems that at this stage, the enactive approach faces a dilemma. In a nutshell, if the enactive account of bodily experiences is cast in terms of spatial know-how, then it is false, because spatial know-how is separable from bodily experiences. If the enactive account is not cast in terms of spatial know-how, then it is unclear what type of sensorimotor expectations could provide both the exteroceptive and the interoceptive spatial content of any kind of tactile experiences, including instantaneous passive touch. Let me briefly develop the two horns of the dilemma.

On the first horn of the dilemma, proponents of the enactive approach characterize the spatial content of tactile experiences in terms of spatial know-how relative to the bodily location that has been touched (how to reach it or move it). But we have seen that this enactive account is challenged by a series of empirical results that reveal that, contrary to what the enactive approach predicts, tactile experiences are separable from the perceiver's actions. On the one hand, the possession of accurate spatial know-how does not guarantee that one consciously feels a touch (as in numbsense). Nor does it guarantee that the spatial content of bodily experiences is accurate (as in the case of patient JO and the Rubber Hand Illusion). On the other hand, the lack of spatial know-how guarantees neither that one lacks bodily experience, nor that one's bodily experience lacks spatial content (as in the case of patient KE). These results are difficult for the enactive approach to accommodate if it aims to explain tactile experiences in terms of the procedural knowledge involved in reaching, pointing, or scratching one's body. What we feel is not determined by what we do, nor by what we know how to do.

On the second horn of the dilemma, proponents of the enactive approach discount spatial know-how involved in such movements as irrelevant for an enactive account of bodily experiences. On this view, they claim to be interested in what grounds perceptual experiences, not in what they are for.[18] Hence, they reject these movements as mere "practical consequences" of bodily experiences. The link between perception and such types of action is said to be only instrumental. It thus is no surprise if this type of spatial know-how is independent and dissociable from bodily experiences. However, if proponents of the enactive view want to argue that spatial know-how does not involve the right kind of sensorimotor knowledge for bodily experiences, then they have to face several difficulties.

First, they must explain why this type of relationship between perception and action is merely instrumental, rather than constitutive. Why couldn't perceptual experiences be grounded in what they are for? It is unclear how to draw the line between the "right" type and the "wrong" type of procedural knowledge. As already mentioned, an enactivist may reject reaching and pointing movements because they do not affect tactile content in the same way that turning one's head affects visual content. This may be true of reaching, pointing, and grasping movements, but not of scratching, where there is a clearly corresponding sensory consequence of one's movements (that is, if one scratches where one has been bitten, the stimulation decreases). Yet, Merleau-Ponty's patient Schneider had no conscious experience of the itching, although he could accurately scratch his leg. If this type of sensorimotor contingency is not included in the set of patterns of interdependence between stimulation and movement that constitutes itching experiences, then it is hard to see what would be. On the other hand, it is not clear that enactivism can relinquish procedural knowledge that cannot be articulated in sensorimotor terms. Consider the experience of a flash of light. What is the enactive account for the *temporal* content of the visual experience (that is, experience of

[18] For this line of argument see, for instance, Noë, "Vision without Representation," in Gangopadhyay, Michael Madary, and Finn Spicer, eds., *Perception, Action, and Consciousness: Sensorimotor Dynamics and Two Visual Systems* (New York: Oxford, 2010), pp. 245–56.

the brevity of the flash)? There is no corresponding sensory expectation of how one's movements would affect the flash of light, given that the flash is already in the past. Furthermore, one cannot appeal to a counterfactual of the type, "If the visual object were still here and if I did *x*, then it would have the visual consequence *y*." Indeed, this would not account for the brevity of the visual stimulus. Hence, enactive explanations must sometimes accept procedural knowledge with no possible modification of the sensory signal. And if this is true for temporal content, why not for spatial content as well? Especially since the enactivist is left empty handed with no sensorimotor account of the spatial content of bodily experiences. As I said in the beginning, it is unclear what sensorimotor contingencies are supposed to be in the case of passive touch, especially if the touch is very brief. One cannot simply highlight the importance of action for bodily experiences; one has to give a detailed account of the specific set of noninstrumental sensorimotor expectations that constitute bodily experiences. If the type of procedural knowledge that I have offered is not viewed as relevant for the enactive approach, then I invite the proponents of such an approach to provide an alternative, which should be empirically testable.

However, that might be an impossible task. As we have seen, the discounting of reaching, pointing, grasping, and scratching movements presupposes a distinction between two types of procedural knowledge (constitutive and instrumental), and thus a distinction between two action systems, one of which is constitutively linked to perceptual experiences and another which is not. If there are such dual sensorimotor systems, it should be possible to dissociate them, but one may wonder whether one sensorimotor system can be impaired while the other is preserved. Alternatively, the enactivist has to postulate that there is a unique action system, but that it exploits two distinct sets of spatial information about the location of the sensation, one of which is identical to the spatial content of bodily experiences.[19] For instance, in the RHI, the location of the hand that was stroked would be encoded as two centimeters farther for one type of action than for another. Thus, the spatial information exploited by the motor system would be internally inconsistent. This seems highly unlikely. As far as I know, there is no empirical validation of such a dissociation within the motor system. Any sensorimotor account of the specific location of a sensation must involve an action that pinpoints that location in the sensorimotor system. And it is unlikely that there can be two distinct sensorimotor systems, or even two distinct sets of spatial information exploited by the sensorimotor system.

---

[19] We know that the same spatial information can be presented within two distinct frames of reference. But this is different from having two sets of spatial information.

To summarize, proponents of the enactive approach either claim that bodily experiences are determined by spatial know-how involved in reaching and pointing movements, or they do not make such a claim. If they do, then the most plausible version of the enactive account of bodily experiences is subject to the concerns I have raised in this paper. If they do not, then the burden of proof is on the enactive side to distinguish between sensorimotor knowledge that does constitute bodily experience and sensorimotor knowledge that does not. Further, enactivists must propose an alternative sensorimotor account of bodily experiences, which assumes the existence either of two sensorimotor systems or of two distinct sets of spatial information exploited by the sensorimotor system. The enactive approach to bodily experiences thus suffers from serious flaws. And if the enactive approach does not work for bodily experiences, which at first sight seem the less controversial example, then it is even less likely to work for visual experiences.

It has been suggested that vision first evolved not to provide conscious perceptual experiences, but to provide distal sensory control of movements. It was only later in evolution, with the emergence of increasingly complex behaviors, that vision came to provide internal models of the world, stored in memory and accessible to other cognitive systems. Perception and action require different transformations of the visual signal. For instance, action requires on-line adjustments, whereas perception, which allows us to recognize objects, is concerned with enduring visual properties. In addition, action requires the object one acts upon to be localized relative to oneself within an egocentric frame of reference, whereas there is no such requirement for perception, which is able to make allocentric judgments about the relative locations of two objects. These distinctions result in the evolution of two visual systems, which interact most of the time but are dissociable.

Similarly, it seems that bodily information processing first evolved for use in action. What is needed for action is different in many respects from what is needed for bodily experiences. As such, the same bodily information is encoded in two different formats, and only one of them is linked to action. For instance, what is needed for action is constantly updated bodily information such that from moment to moment one knows the posture of one's body parts. By contrast, we have seen in the RHI study that participants still experience their hand as closer to the rubber hand after having moved it. This shows that their bodily experience was not updated by proprioceptive information. These findings, among others, suggest that there are two functionally defined somatosensory systems, which interact most of the time but are dissociable. One may further suggest that they lead to two distinct, functionally defined representations of the body: the body schema for action (that is, information about the body necessary for movement, such as posture, limb size, and strength) and the body image for perception (that is, judgment of one's own bodily properties). The dissociations in touch and proprioception show that vision is not an exception in the architecture of the mind. Other modalities, including touch and proprioception, use dual coding based on the functional distinction between perception and action. The brain encodes information in distinct formats, and the way the brain uses information determines the way it encodes it.

FRÉDÉRIQUE DE VIGNEMONT

Institut Jean-Nicod

COMMENTS AND CRITICISM

METHODOLOGY AND THE NATURE OF KNOWING HOW

In their delightfully provocative paper, "Knowing How,"[1] Jason Stanley and Timothy Williamson (S&W) reject Gilbert Ryle's thesis[2] that "there is a fundamental distinction between knowledge-how and knowledge-that" (411). Indeed, they seem to regard that widely accepted thesis as nothing but a "philosophical prejudice" (431). They offer an ingenious linguistic argument for the sharply different view that "knowledge-how is simply a species of knowledge-that" (411).

S&W's argument starts from a claim about the syntax of English *ascriptions* of knowledge-how. Their paradigm ascription is:

(P) Hannah knows how to ride a bicycle. (417)

They claim that ascriptions like (P) contain embedded questions. They go on to argue that, according to the "standardly accepted semantics," such ascriptions attribute propositional knowledge (420). From this they arrive at their conclusion—since aptly called "Intellectualist" by Stanley[3]—that knowledge-how is a species of knowledge-that.

Why should we care about knowledge-how? From a naturalistic perspective, we should care only if knowledge-how is a feature of mental reality that is pertinent to psychological explanations. And philosophers have thought that it is, as S&W note. They discuss two "representative cases" (441–44): David Lewis's appeal[4] to knowledge-how in responding to Frank Jackson's famous "knowledge argument" about Mary; and my appeal[5] to it in discussing linguistic competence.

I do indeed appeal to it, saying that linguistic competence is mere knowledge-how. However, I have always preferred to say that the competence is simply a skill or ability.[6] Still, I am strongly committed

---

[1] Jason Stanley and Timothy Williamson, "Knowing How," this JOURNAL, XCVIII, 8 (August 2001): 411–44 All unidentified references in the text are to this work.

[2] Gilbert Ryle, *The Concept of Mind* (Chicago: University Press, 1949), and "Knowing How and Knowing That," in *Gilbert Ryle. Collected Papers*, vol. 2 (New York: Barnes and Noble, 1971), pp. 212–25.

[3] Stanley, "Knowing (How)," *Noûs*, XLV, 2 (June 2011): 207–38.

[4] David Lewis, "What Experience Teaches," in William G. Lycan, ed., *Mind and Cognition: An Anthology* (Malden, MA. Blackwell, 1990), pp. 499–519.

[5] Michael Devitt, *Coming to Our Senses* (New York: Cambridge, 1996).

[6] *Ibid.*, pp. 22–28, 52; Devitt, *Designation* (New York: Columbia, 1981), pp. 92–110, and *Realism and Truth*, 2nd ed. (Princeton: University Press, 1997), pp. 272–75; Devitt and

0022-362X/11/0804/205–218

to the competence *not* being knowledge-that. So, if we must follow the folk in talking of linguistic competence as "knowledge" at all, then it looks as if I must accept that the competence is knowledge-how. I think that it would be better if we did not follow the folk in this talk,[7] but there is no doubt that it is tempting to do so. So, I clearly have a particular interest in rejecting S&W's view that knowledge-how entails knowledge-that.

The distinction between knowledge-how and knowledge-that is, of course, a folk-psychological one captured in English using the very loose term 'know'. My take on the distinction has always been along the following rather Rylean lines.[8] Knowledge-that is essentially cognitive and propositional. So if a person knows that *R* is a rule of arithmetic she knows a proposition. Knowledge-how, in contrast, is in the same family as skills, abilities, and capacities. Sometimes it is entirely cognitive; for example, knowing how to play chess. Other times, it may be hardly cognitive at all; for example, knowing how to swim or ride a bicycle.[9] Sometimes, knowledge-how may involve knowledge-that; for example, chess know-how may involve knowing that the rules of chess are such and such. Other times it is, as the folk would say, "mere know-how" and *prima facie* does not involve any propositional knowledge at all; for example, knowing how to swim.

Paul Snowdon[10] and John Bengson, Marc Moffett, and Jennifer Wright[11] have made me think that this account is probably *too* Rylean. There is evidence that the folk count a person who can give a full description of an activity as knowing how to perform it even though the person has no ability to perform it. So, perhaps *one* kind of knowledge-how is knowledge-that. But I still want to maintain that another kind is not. There is a common kind of knowledge-how that a person can have simply on the basis of having the ability to perform an activity:[12]

Kim Sterelny, *Language and Reality: An Introduction to the Philosophy of Language*, 2nd ed. (Cambridge: MIT, 1999), pp. 187–90.

[7] Devitt, *Ignorance of Language* (New York: Oxford, 2006), p. 5n5.

[8] Devitt and Sterelny, "What's Wrong with 'The Right View'," in James E. Tomberlin, ed., *Philosophical Perspectives 3 Philosophy of Mind and Action Theory* (Atascadero, CA: Ridgeview, 1989), pp. 497–531, and *Language and Reality*, pp. 174–75; Devitt, *Ignorance of Language*, pp. 46–47, 50

[9] As these examples illustrate, the concern here is with knowing how *to A* where *A* is some activity.

[10] Paul Snowdon, "Knowing How and Knowing That: A Distinction Reconsidered," *Proceedings of the Aristotelian Society*, n.s., CIV (2004) 1–29.

[11] John Bengson, Marc Moffett, and Jennifer Wright, "The Folk on Knowing How," *Philosophical Studies*, CXLII, 3 (February 2009). 387–401.

[12] Bengson, Moffett, and Wright have conducted an experiment that leads them to doubt this. See *ibid.*, p. 397. I am not persuaded by the experiment but will not attempt to argue the matter here.

the knowledge may be *mere* knowledge-how. That is the kind that Ryle had in mind. If there really is such a kind then, contrary to S&W, not all knowledge-how is knowledge-that.

My paper has two aims. The first is to argue that the Intellectualist thesis is false. My argument starts in section I by rejecting the alleged intuitive support for the thesis. But the core of my case against Intellectualism comes in sections II and III from something ignored by S&W, the *science* of knowledge-how. Finally, in section IV, I look critically at S&W's particular Intellectualist thesis. If my argument against Intellectualism is right then there must be something wrong with the semantic theory that leads S&W to it, but I shall make no serious attempt to say what.[13] My second aim is to argue that S&W's methodology of deriving a theory of knowledge-how from a linguistic theory of ascriptions, without any attention at all to the science of knowledge-how, is deeply misguided. That is the concern of the concluding section V.

I. INTUITIONS

S&W rightly think that their semantic theory of knowledge-how ascriptions like (P) would gain support if analogous ascriptions were found to ascribe relations between persons and propositions. They claim that analogues "clearly" do ascribe such relations (419, 422). They offer no reasons independent of their linguistic theory for these claims; the claims seem to rest solely on S&W's intuitions.[14] I shall start by considering these intuitions. I do not share them. With almost all the analogues, it is far from clear that they ascribe relations to propositions. And with some it seems to me fairly clear that they do *not* ascribe such relations, just as it does that knowledge-how ascriptions do not.

(a) S&W first consider a group of ascriptions that are like (P) but ascribe knowledge-*where*, knowledge-*whom*, knowledge-*which*, or knowledge-*why*; for example, 'Hannah knows where to find a nickel' (418). For a person to know why, she must have reasons, and so knowledge-why does indeed seem to be propositional. But what about the others? It is helpful to think of animals. The foraging desert ant wanders all over the place until it finds food and then always heads straight back to its nest.[15] On the strength of this competence, we feel no qualms about saying that it "knows where to find its nest." But to attribute any propositional attitudes to the ant simply on the strength of that competence seems like soft-minded anthropomorphism. Cognitive ethologists might lead us in that direction, of course—on which more in section III—but we do not seem to get knowledge-that *for nothing* with knowledge-where.[16]

---

[13] Ian Rumfitt, "Savoir Faire," this JOURNAL, C, 3 (March 2003): 158–66, however, does make a serious attempt, claiming that the semantics does not hold up when we consider other languages. Thus, although some ascriptions of knowledge-how in French may fit S&W's analysis, those ascribing knowing how to swim, ride a bicycle, and the like do not. These ascriptions seem to attribute a relation to an activity, not to a proposition, just as those who emphasize the distinction between knowledge-how and knowledge-that would hope. Stanley responds to this and other cross-linguistic challenges in a follow-up paper, *op. cit.*

[14] Snowdon expresses some similar ones, *op. cit.*

(b) S&W next consider ascriptions that are like (P) and the first group except that they are tensed and refer to another person; for example, 'Hannah knows how Bill rides a bicycle' (418). Again none of these ascriptions, except the ascription of knowledge-why, seem "clearly" to ascribe propositions. Thus mightn't Hannah demonstrate this knowledge-how *simply* by imitating Bill riding a bicycle? And consider the western scrub jay. One of these, *X*, may observe another, *Y*, caching food, and later raid the cache.[17] We are inclined to say that *X* knows where *Y* cached the food. Ethologists think that these jays are pretty smart, but *must* they say that this knowledge-where is propositional?

(c) Finally, S&W consider ascriptions of *learning*-how, *recalling*-how, *asking*-how, *wondering*-how, *being certain*-how, *indicating*-how, and *seeing*-how. Asking-how seems propositional, but none of the others do—certainly not *clearly* propositional. Chicks learn how to fly and babies see how to walk. This does not seem like grasping propositions. And the psychological evidence of implicit learning, mentioned briefly in section III, counts heavily against the view that learning-how must be propositional.

In sum, I suggest that consideration of the analogous ascriptions does not give intuitive support to S&W's view of knowledge-how. Rather, it suggests that S&W should take another look at their semantic theory.

But all of this is "mere intuition." I turn now to more serious objections to S&W's Intellectualism. The most serious of all comes from the psychology of skills.

II. THE PSYCHOLOGY OF SKILLS[18]

The folk distinction between knowledge-that and knowledge-how is commonly thought to be the same as the psychological one between

---

[15] Described in Charles R. Gallistel, *The Organization of Learning* (Cambridge: MIT, 1990).

[16] Cf. Stanley's claim that knowledge-where is "definable" in terms of propositional knowledge, *op. cit.*

[17] Nicola Clayton, Nathan Emery, and Anthony Dickinson, "The Rationality of Animal Memory: Complex Caching Strategies of Western Scrub Jays," in Susan Hurley and Matthew Nudds, eds., *Rational Animals?* (New York: Oxford, 2006), pp. 197–216.

[18] I draw on a much more detailed discussion in *Ignorance of Language*, pp. 210–20.

"declarative" and "procedural" knowledge. That distinction originated in artificial intelligence but is widely acknowledged, frequently applied, and very important in psychology. Thus, John Anderson, a leading cognitive psychologist, writes:

> The distinction between *knowing that* and *knowing how* is fundamental to modern cognitive psychology. In the former, what is known is called *declarative knowledge*; in the latter, what is known is called *procedural knowledge*.[19]

Psychologists describe the distinction, rather inadequately, along the following lines: where declarative knowledge is explicit, accessible to consciousness, and conceptual, procedural knowledge is implicit, inaccessible to consciousness, and subconceptual. Although declarative knowledge may play a role in learning a skill, there is consensus that the skill itself is a piece of procedural knowledge.

This psychologists' distinction is related to two others. First, there is a distinction between explicit (or declarative) and implicit (or procedural) *memory*. Declarative knowledge involves explicit memory; procedural knowledge involves implicit memory. Explicit memory holds factual knowledge such as that Washington DC is the capital of America, while implicit memory holds rules that govern processes, "routinized skills...priming, and classical and operant conditioning."[20] Second, there is a distinction between explicit and implicit *learning*. Explicit learning is a "top-down" process that starts from declarative knowledge. Consider, for example, learning to change gears in a stick-shift car by starting with instructions like: "First, take your foot off the accelerator, then disengage the clutch." In contrast, implicit learning is a "bottom-up" process: we observe, practice, and "just pick the skill up." A. S. Reber defines implicit learning as follows: "the capacity to pick up information about complex stimulus displays largely without awareness of either the process or the products of learning."[21] She notes that systems of implicit learning "are rather easily simulated by connectionist architectures."[22] There is much evidence that a lot of skill learning is implicit; see, for example, the evidence cited by Ron Sun, Edward Merrill, and Todd Peterson that "individuals may learn complex skills without first obtaining a large amount of explicit declarative knowledge...and without being able to verbalize the rules they use."[23]

---

[19] John R. Anderson, *Cognitive Psychology and Its Implications* (San Francisco: W. H. Freeman, 1980), p. 223.

[20] David F Bjorklund, Wolfgang Sneider, and Carlos Hernandez Blasi, "Memory," in Lynn Nadel, ed., *Encyclopedia of Cognitive Science*, vol. 2 (New York: Nature Publishing Group, 2003), pp. 1059–65, at p. 1059.

[21] A. S. Reber, "Implicit Learning," in *ibid*, pp. 486–91, at p. 486.

[22] *Ibid.*, p. 487.

What is declarative knowledge? The key thing to note for our purposes is that there is consensus in psychology that it involves a *conscious representation* of what is known. Thus, psychologists think that a subject has declarative knowledge of the processing rules for a task only if she consciously represents the rules. So the person who has declarative knowledge that *R* is a rule of arithmetic must represent that fact in her central processor. So, if we adopt the popular, and in my view correct, representational theory of the mind (RTM), declarative knowledge can indeed be identified with the folk's knowledge-that.

Before considering the nature of procedural knowledge, it is helpful to mention another distinction that I have emphasized elsewhere.[24] It comes largely from computer science. It is the distinction between processing rules that govern by being represented and applied and those that govern by being simply embodied, without being represented. This is a distinction between two ways in which certain processing rules might be real in an object, two ways in which the rules might be embodied in it. Neither of these ways should be confused with a situation where an object simply behaves *as if* it is governed by those processing rules. For that situation is compatible with those rules *not* being embodied in the object at all.

A simple old-fashioned mechanical calculator provides a nice example of something governed by rules that are embodied without being represented. When the calculator adds it goes through a mechanical process that is governed by the rules of an algorithm for addition. But the rules are "hardwired," not represented in the calculator. In contrast, the operations of a contemporary general-purpose computer are partly governed by rules of a program that are represented in its RAM and applied. Yet those rules can govern the operations of the computer only because there are other, unrepresented rules built into its hardware which enable the represented rules to govern. And, note an important generalization: any processing rule that governs the behavior of one object by being represented and applied could govern that of another by being embodied without being represented.

What is procedural knowledge, and hence, what are skills? It is agreed that this knowledge is *not* declarative but, beyond that, we have a long way to go in answering this question.[25] The psychological literature reveals a range of interesting ideas but no rational basis at this time for a sweeping acceptance or rejection of the ideas of one or other theoretical camp. What about the representation of rules? It would be nice if procedural knowledge consisted in embodied but unrepresented rules of the sort just mentioned. That would yield the desired sharp distinction between procedural and declarative knowledge of rules. Certainly, there is no persuasive evidence that skills *do* involve representations of the governing rules. But neither is there decisive evidence that the skills do *not* involve these representations. Still, the literature on motor skills,[26] on dynamical systems theories,[27] on the Gibson-inspired ecological theories,[28] on connectionist theories,[29] on instance theories,[30] on implicit learning,[31] and on the implementation of skills[32] counts heavily against the idea that skills do involve those representations. The only support for that idea may come from production-systems theories which seem committed to representations of rules in procedural knowledge, albeit representations of a different sort from those involved in declarative knowledge.[33] But, I argue, there are reasons for doubting the appropriateness of this apparent commitment.[34]

---

[23] Ron Sun, Edward Merrill, and Todd Peterson, "From Implicit Skills to Explicit Knowledge: A Bottom-Up Model of Skill Learning," in *Cognitive Science*, xxv, 2 (March–April 2001): 203–44, at p. 207.

[24] Devitt, *Ignorance of Language*, pp. 45–52.

---

[25] Daniel L. Schacter, "Implicit versus Explicit Memory," in Robert A. Wilson and Frank C. Keil, eds., *The MIT Encyclopedia of the Cognitive Sciences* (Cambridge. MIT, 1999), pp. 394–95, at p. 395; Sun, "Connectionist Implementationalism and Hybrid Systems," in Nadel, ed., *op. cit.*, vol. 1, pp. 697–703, at p. 698.

[26] See, for example, Liana E. Brown and David A. Rosenbaum, "Motor Control: Models," in *ibid.*, vol. 3, pp. 127–33; Mark Mon-Williams, James R. Tresilian, and John P. Wann, "Motor Control and Learning," in *ibid*, pp. 121–26; Daniel M. Wolpert and Zoubin Ghahramani, "Motor Learning Models," in *ibid*, pp. 138–42.

[27] See, for example, J. A. Scott Kelso, *Dynamic Patterns: The Self-Organization of Brain and Behavior* (Cambridge: MIT, 1995); Tim van Gelder, "Dynamic Approaches to Cognition," in Wilson and Keil, eds., *op. cit.*, pp 244–46; James Garson, "Dynamical Systems, Philosophical Issues about," in Nadel, ed., *op. cit.*, vol. 1, pp. 1033–39; Richard A. Carlson, "Skill Learning," in Nadel, ed., *op. cit.*, vol 4, pp. 36–42.

[28] See, for example, Carol A. Fowler and Michael T. Turvey, "Skill Acquisition: An Event Approach with Special Reference to Searching for the Optimum of a Function of Several Variables," in George E. Stelmach, ed., *Information Processing and Motor Control* (New York: Academic, 1978), pp. 2–40; Peter Noble Kugler and Turvey, *Information, Natural Laws, and the Self-Assembly of Rhythmic Movement* (Hillsdale, NJ: Laurence Erlbaum, 1987); Karl M. Newell, "Motor Skills," in Larry R. Squire, ed., *Encyclopedia of Learning and Memory* (New York: Macmillan, 1996), pp. 441–43.

[29] See, for example, David E. Rumelhart and James L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (Cambridge: MIT, 1986); Michael E. J. Masson, "Cognitive Theories of Skill Acquisition," *Human Movement Science*, ix, 3–5 (September 1990): 221–39; Sun, Merrill, and Peterson, *op. cit.*

[30] See, for example, Gordon D. Logan, "Toward an Instance Theory of Automatization," *Psychological Review*, xcv, 4 (October 1988): 492–527.

[31] See, for example, Robert C. Mathews, Ray R. Buss, William B. Stanley, and Roberta Chinn, "Analysis of Individual Learning Curves in a Concept Discovery Task. Relations among Task Performance, Verbalizable Knowledge, and Hypothesis Revision Strategies," *The Quarterly Journal of Experimental Psychology*, xlA (1988): 135–65; Mathews et al., "Role of Implicit and Explicit Processes in Learning from Examples: A Synergistic Effect," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, xv, 6 (1989): 1083–100; Reber, "Implicit Learning and Tacit Knowledge," *Journal of Experimental Psychology: General*, cxviii, 3 (September 1989): 219–35, and "Implicit Learning"; William B. Stanley et al., "Insight without Awareness: On the Interaction of Verbalization, Instruction, and Practice in a Simulated Process Control Task," *The Quarterly*

In sum, the psychological distinction between declarative and procedural knowledge, and the related distinctions about memory and learning, are well established in empirical science. As one researcher says, the evidence for them "lies in experimental data that elucidate various dissociations and differences in performance under different conditions."[35] And all of this is evidence for the nature of, near enough, the folk distinction between knowledge-that and knowledge-how.

I say "near enough" because I think that the psychologists would have done better to identify their procedural knowledge with one common kind of the folk's knowledge-how, *mere* knowledge-how. For, as we noted, it is that Rylean kind of knowledge-how that is thought not to involve knowledge-that. Still, we need not fuss about this: mere knowledge-how is still knowledge-how. So, if the psychologists are right and procedural knowledge does not involve declarative knowledge, then declarative knowledge is not essential to knowledge-how.

This psychology is very bad news for S&W. Knowledge-that is declarative knowledge. So, psychology shows that procedural knowledge, hence knowledge-how, does not require knowledge-that, precisely

---

Journal of Experimental Psychology, XLIA, 3 (1989): 553–77; Sun, Merrill, and Peterson, *op. cit.*; Axel Cleeremans, "Implicit Learning Models," in Nadel, ed., *op. cit.*, vol. 2, pp. 491–99.

[32] See, for example, Atsushi Yamadori et al., "Neurological Basis of Skill Learning," *Cognitive Brain Research*, v, 1–2 (December 1996): 39–54; Michael I. Posner, Gregory J. DiGirolamo, and Diego Fernandez-Duque, "Brain Mechanisms of Cognitive Skills," *Consciousness and Cognition*, VI, 2–3 (June 1997): 267–90.

[33] See, for example, Anderson, *The Architecture of Cognition* (Cambridge: Harvard, 1983), and *Rules of the Mind* (Hillsdale, NJ: Lawrence Erlbaum, 1993); John E. Laird, Allen Newell, and Paul S. Rosenbloom, "Soar: An Architect for General Intelligence," *Artificial Intelligence*, XXXIII, 1 (1987): 1–64, Mark K. Singley and Anderson, *The Transfer of Cognitive Skill* (Cambridge: Harvard, 1989); Anderson and Christian Lebiere, *The Atomic Components of Thought* (Mahwah, NJ: Lawrence Erlbaum, 1998); Masson, *op. cit.*; Lebiere, "ACT," in Nadel, ed , *op. cit.*, vol. 1, pp. 8–14; Gary Jones, "Production Systems and Rule-Based Inference," in Nadel, ed., *op. cit.*, vol. 3, pp. 741–47.

[34] Devitt, *Ignorance of Language*, pp. 215–17.

[35] Sun, *op. cit* , p. 698. See also Schacter, *op. cit.*, p. 394; the many results cited by Sun, Merrill, and Peterson, *op. cit.*, p. 207; Cleeremans, *op. cit.*, p. 492; Neil W. Mulligan, "Memory. Implicit versus Explicit," in Nadel, ed., *op. cit.*, vol. 2, pp. 1114–20, especially pp. 1115–17; Reber, *op. cit.*, p. 491.

what S&W's Intellectualism claims it does require. Despite disagreement or uncertainty on many other issues, psychologists speak with one voice on this one. Even production-system theories, which seem to posit representations of processing rules and hence seem to have the most "intellectualized" picture of procedural knowledge, distinguish this knowledge sharply from declarative knowledge. Psychology presents a picture of procedural knowledge as constituted somehow or other by embodied, probably unrepresented rules that are inaccessible to consciousness. It is thus quite different from declarative knowledge which consists of representations that are available to consciousness.

III. COGNITIVE ETHOLOGY

The distinction between procedural and declarative knowledge is also central to cognitive ethology: "Declarative knowledge is 'knowing that' whereas procedural knowledge is 'knowing how,' or knowing what to do, as in a stimulus-response connection."[36] It is obvious that animals, from insects to primates, have a vast range of skills, many of them innate.[37] The very difficult challenge in explaining the nature of each of these skills is to find the right place between a "rash anthropomorphism" that attributes declarative knowledge to the organism on little evidence, and a crass behaviorism that implausibly attributes only the simplest forms of procedural knowledge.[38] Ethologists (and comparative psychologists) take up this challenge by seeking evidence that forces them away from strictly behaviorist explanations. And they have found plenty, leading them to posit surprisingly rich cognitive lives; for example, evidence in navigation by insects,[39] in caching strategies by western scrub jays,[40] and in language learning by bottlenosed dolphins.[41] But I see no sign in the literature of retreat from the received view that procedural knowledge is quite distinct from declarative knowledge.

---

[36] Sara J. Shettleworth, *Cognition, Evolution, and Behavior* (New York: Oxford, 1998), p 5, see also David McFarland, "Defining Motivation and Cognition in Animals," *International Studies of Philosophy of Science*, v, 2 (1991): 153–70.

[37] Sometimes, the term 'skill' is restricted to something learnt; Carlson, *op. cit.*, p. 36. This is not wise verbal legislation. Something with the nature of a skill that is, as a matter of fact, learnt, could, in another possible world, be innate. We need a theory of that nature whatever its source.

[38] Gordon M. Burghardt, "Animal Awareness. Current Perspectives and Historical Perspectives," *American Psychologist*, XL, 8 (1985)· 905–19, at p. 905.

[39] Described in Gallistel, *op. cit.*

[40] Clayton, Emery, and Dickinson, *op. cit.*

[41] Louis M. Herman, "Exploring the Cognitive World of the Bottlenosed Dolphin," in Marc Bekoff, Colin Allen, and Gordon M. Burghardt, eds., *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition* (Cambridge: MIT, 2002), pp. 275–83, and "Intelligence and Rational Behavior in the Bottlenosed Dolphin," in Hurley and Nudds, eds., *op. cit.*, pp. 439–67.

At one point S&W briefly address an objection that arises from considering animals:

> in certain situations, we smoothly ascribe knowledge-how to animals.... One might think that nonhuman animals are not sufficiently conceptually sophisticated enough to possess propositional knowledge. But this objection is a nonstarter. For in similar scenarios, we just as smoothly ascribe propositional knowledge to nonhuman animals. (438–39)[42]

This swift response will not do. Cognitive ethologists and comparative psychologists do indeed ascribe *some* propositional attitudes to *some* animals, but they are mighty cautious about doing so.[43] They are not similarly cautious about ascribing knowledge-how. Indeed, it is simply taken for granted that nonhuman animals have a great deal of knowledge-how or procedural knowledge: "there seems little doubt that nonhuman animals have varying degrees of procedural knowledge."[44] Even traditional ethologists and behavioral scientists accept that they have. And why wouldn't they, since the results of conditioned and associative learning count as procedural knowledge?

> An explanation of how an animal *knows how* might simply involve a stimulus-response connection or the fruit of some other form of associative learning, or it might involve the attribution of something more complex, such as the internal representation of the external environment—a cognitive map of the world in which it lives.[45]

So there is no caution about saying, for example, that the earlier-mentioned ant knows how to get to the nest,[46] and the earlier-mentioned jay knows how to raid a cache; about saying that a honey bee knows how to communicate the direction and distance of a food source using a "waggle dance"; and so on through much of the behavior of animals. Yet I doubt that S&W could find a single ethologist or psychologist who would be prepared to ascribe to these animals, particularly to insects, the propositional knowledge that—we shall see in the next section—S&W's account of knowledge-how requires.[47]

---

[42] Consider the following also: "In those cases where attributions of know-how are scientifically indispensible, we are comfortable with the corresponding attributions of propositional knowledge... " Bengson, Moffett, and Wright, *op. cit.*, p 399.

[43] For helpful discussions, see Burghardt, *op. cit*; Shettleworth, *op cit.*; Bernd Heinrich, "Raven Consciousness," in Bekoff, Allen, and Burghardt, eds., *op. cit.*, pp. 47–52.

[44] Niall Shank, *Animals and Science: A Guide to the Debates* (Santa Barbara, CA. ABC-CLIO, 2002), p. 211.

[45] *Ibid*

[46] On the role of procedural knowledge in ant navigation, see Markus Knaden, Christina Lange, and Rüdiger Wehner, "The Importance of Procedural Knowledge in Desert-Ant Navigation," *Current Biology*, xvi, 21 (November 2006): R916–17.

[47] Thanks to Colin Allen and Robert Lurz for advice on this paragraph.

In sum, a nonpropositional view of knowledge-how is not just philosophical prejudice or even just folk theory: it seems to be entrenched in psychology and cognitive ethology. We have very good reasons to suppose that there are nonpropositional states of knowledge-how, constituted probably by embodied but unrepresented rules, states that are central to explaining many behaviors. Could *any* semantic theory be strong enough to overthrow this? I think that is most unlikely. I love some semantic theories, myself, but I do not think that any one of them is well-enough established to have this metaphysical clout. S&W's semantic theory must be mistaken.

Perhaps not. In supposing that S&W's semantics of knowledge-how ascriptions is inconsistent with the psychology of procedural knowledge, I am presuming that these ascriptions sometimes *succeed* in describing the reality of procedural knowledge. But perhaps they never do succeed. Perhaps S&W should be seen not as challenging psychological theories of procedural knowledge, but only the view that this knowledge is ever knowledge-how, that it is a kind of knowledge-how. So S&W have shown us that the procedural knowledge that psychologists and ethologists think of as knowledge-how, and much of which the folk would also surely think of as knowledge-how, is not really knowledge-how. That indeed would be a bold claim to rest on a semantic theory.

These last two sections have presented the main case against S&W's Intellectualism. In the next section I will add a little to this case by looking at the details of their view.

IV. STANLEY AND WILLIAMSON'S INTELLECTUALISM

S&W's view that all knowledge-how is knowledge-that is not only at odds with psychology and ethology; it is highly implausible. Stephen Schiffer has given a brief but neat demonstration of this with his examples of the young Mozart and Clyde the chicken-sexer.[48] I shall attempt to add to Schiffer's case.

Consider what S&W tell us about their paradigm: (P) "is true relative to a context *c* if and only if there is some contextually relevant way *w* such that Hannah stands in the knowledge-that relation to the Russellian proposition that *w* is a way for Hannah to ride a bicycle, and Hannah entertains this proposition under a practical mode of presentation" (430). Schiffer notes that "we are not given the slightest clue as to what 'practical modes of presentation' are like."[49] Set that aside. According to S&W, for (P) to be true Hannah has to have the thought that *w* is a way for Hannah to ride a bicycle. This requires, of course, that Hannah has the *general* concept of ways of riding a bicycle, which she may well have. But it also requires that she has a *singular* concept that *identifies* a type of action *w* which is, as a matter of fact, a way of riding a bicycle. Is it plausible that she, indeed *every* bicycle rider, has identified such an action? S&W claim, reasonably enough, that this identification need not be descriptive: it can be indexical (433). But how plausible is it that riders even have an indexical concept of such an action?

---

[48] Stephen Schiffer, "Amazing Knowledge," this JOURNAL, XCIX, 4 (April 2002), 200–02.

[49] *Ibid.*, p. 201.

Take a favorite example of mine.[50] Outfielders in professional baseball know how to catch fly balls. An essential part of this is being at the right place when the ball descends to catch height. An experiment showed that skilled fielders "ran at a speed that kept the acceleration of the tangent of the angle of elevation of gaze to the ball at 0."[51] So, according to S&W, fielders have a singular concept of this type of movement. Fielders do not have to think of the movement in those terms, of course, but they still have to have identified it. It does not seem plausible that all fielders can manage this.

It gets worse. Some people know how to think rationally. According to S&W they must have a singular concept that identifies something that is, as a matter of fact, a way of thinking rationally. Does *anyone* have such a concept? And *must* everyone who thinks rationally have the general concept of thinking rationally, or even the concept of thinking?

Let us apply S&W's view to the case of linguistic knowledge. Hannah knows how to speak a language *L*. So, she knows how to use a sound of *L* to express a thought with the meaning that the sound has in *L* in the context of utterance. Once again, does *anyone* have a singular concept of something that is a way of doing this? Must *everyone* who can do it have the sophisticated general concept of a way of doing it? To suppose so is to fly in the face of developmental evidence that the capacity to think about one's language does not normally come until middle childhood, well after linguistic competence.[52] So, it is very likely that Hannah, in early childhood, did not even have the concept of L and so *could not* have had the sophisticated concept. Yet she already knew how to speak L.

---

[50] Devitt, *Ignorance of Language*, p. 50.

[51] Peter McLeod and Zoltan Dienes, "Do Fielders Know Where to Go to Catch the Ball or Only How to Get There?" *Journal of Experimental Psychology: Human Perception and Performance*, xxii, 3 (June 1996): 531–43, at p. 531.

[52] See particularly David T. Hakes, *The Development of Metalinguistic Abilities in Children* (New York: Springer-Verlag, 1980); Ellen Bouchard Ryan and George W. Ledger, "Learning to Attend to Sentence Structure: Links Between Metalinguistic Development and Reading," in John Downing and Renate Valtin, eds., *Language Awareness and Learning to Read* (New York: Springer-Verlag, 1984), pp. 149–71; Ellen Bialystok and Ellen Bouchard Ryan, "A Metacognitive Framework for the Development of First and Second Language Skills," in Donna-Lynn Forrest-Pressley, G. E. MacKinnon, and Thomas Gary Waller, eds., *Metacognition, Cognition, and Human Performance* (New York: Academic,

This concludes my case against S&W's Intellectualism. I turn finally to their methodology.

V. METHODOLOGY

On the basis of assumptions about linguistic expressions that concern the psychological world, without any attention to the science of that world, S&W draw a very bold conclusion about the nature of that world. I think that we should always be suspicious of this way of proceeding: our semantic theories should be guided by our theories of the world rather than vice versa. We should follow the methodology of "putting metaphysics first."

Why? *Because we know much more about the way the world is than we do about the semantics of our talk about that world.* I have argued for this, and the methodology it supports, in the course of arguing against the linguistic turn in contemporary philosophy. Similarly, I have argued that we should reject the epistemological turn in modern philosophy because we know much more about the way the world is than we do about how we know about that world. We should approach epistemology and semantics from a metaphysical perspective rather than vice versa. My argument for this position reflects a somewhat Moorean approach. More importantly, it reflects Quinean naturalism: the metaphysics I want to put first is a naturalized one.[58]

Modern philosophy starts its arguments with what seem to be a priori reflections on epistemology. Contemporary philosophy starts its arguments with what seem to be a priori reflections on semantics. S&W's starting point is importantly different: contemporary linguistics. This move to empirical science is a great improvement. Still, their argument proceeds in the wrong direction. Contemporary linguistics is undoubtedly successful, but it is most unlikely that it could refute any well-established view about the nature of the nonlinguistic world, any well-established "metaphysical" views. The view of knowledge-how that S&W claim to refute is, we have seen, such a view.

1985), pp. 207–52; Bialystok, "Factors in the Growth of Linguistic Awareness," *Child Development*, LVII, 2 (April 1986): 498–510; Carson T. Schütze, *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology* (Chicago: University Press, 1996).

[58] Devitt, *Coming to Our Senses*; *Realism and Truth*; and *Putting Metaphysics First: Essays on Metaphysics and Epistemology* (New York: Oxford, 2010).

S&W are not at all bashful about proceeding from linguistics to psychology (see, for example, 431). Furthermore, they are critical of the methodology of their opponents: "Surprisingly...none of Ryle's followers has ever given an explicit syntax and semantics for [ascriptions of knowledge-how], much less one which would give them the interpretations they claim such constructions to have" (432). But there is nothing surprising about this at all! Scientists have typically, perhaps even always, described their domains with little attention to the syntax and semantics of the descriptions they use for this purpose. Why should psychology, whether folk or scientific, be any different? And it is just as well that science has proceeded in this way, else we would have had very little of it. For only recently have we had powerful syntactic theories. And semantic theories are still in their infancy. The real surprise in this debate is provided by S&W themselves. For they offer a theory of knowledge-how on the basis of linguistics without any attention at all to the science of knowledge-how. Linguistics is surely wonderful but it cannot cure wooden legs.

MICHAEL DEVITT

City University of New York Graduate Center

BOOK REVIEWS

Why Some Things Should Not Be for Sale: The Moral Limits of Markets. DEBRA SATZ. New York: Oxford University Press, 2010. xi + 252 p. Cloth $35.00.

I met Robert Nozick once, in December 1999, after my talk on the meaning of life at Boston University. Over dinner, Nozick said he wanted me to understand his rejection of libertarianism. His reasoning would have resonated with Debra Satz. Nozick, barely thirty in the early 1970s, had said that governments had no right to block "capitalist acts between consenting adults" (quoted by Satz, at 174). The view would define him as "the Libertarian."

By 1999, though, Nozick had moved on (as had libertarianism). What changed Nozick's mind? On the one hand, uncompromising respect for liberty arguably entails respecting your right to sell yourself into slavery. It is your life. Yet, Nozick came to believe that, apart from what actions accomplish, there are questions about what actions symbolize. Even if contracting into slavery were a way of exercising freedom, legally enforcing such contracts would not be a way of *standing* for freedom.

The point is, everyone draws a line. Nozick would have agreed with Satz: some things should not be for sale. Whether rights exist may be controversial, but the actual question at hand is whether such rights as we have can be inalienable. (A right to x normally implies that we may not dispose of x without the owner's consent. An *inalienable* right implies that we may not dispose of x even with the owner's consent.) When we regard my kidney and my vote as my property, yet deny that this entitles me to sell them, we are treating my right as inalienable.

So, why not sell everything? Satz considers what gets bought and sold: sex, weapons, pollution, votes, mercenary services, eternal salvation. She disavows sweeping theories about why such transactions are wrong, and indeed does not assume they always are wrong (210), or that wrongness is sufficient grounds for legal prohibition. Making it illegal for children to work might result in children not working, or it might result in children working for criminals.

Four circumstances make for what Satz calls noxious markets. People can be rendered vulnerable by economic desperation, so that their best option is to accept unconscionable terms of trade. Second, traders can be uninformed to a point where it compromises their agency. Third and fourth, transactions can pose an extreme risk for traders, or threaten to undermine background conditions for equal citizenship, thereby

0022-362X/11/0804/219–223

posing a menace to society at large. We might add a fifth category: buyers might simply be better off if a certain product is not brought to market. Blackmail is an industry that invests in producing information that customers pay to make go away. Customers would be better off had that information never been brought to market.

If we shouldn't sell everything, why sell anything? To Satz, knowing where markets should stop requires knowing how far markets should go, and why. Satz takes her cue from Adam Smith. Trade is the essence of cooperation. If (and only if) the rule of law framework is structured well enough will merchants spend their time making their products maximally useful to their fellow citizens. In the process they achieve equality in its most potently liberating form. Satz credits Smith with the insight that in a market, self-owning workers have a right to walk away. This formal right to walk away changes everything—in principle and in historical practice—liberating people from "abject dependence on one powerful person by allowing them to sustain themselves through exchanges with thousands of anonymous and indifferent customers" (42). To Smith, ending servility was "by far the most important" effect of markets. One answer to our question, then, is that transacting should stop at a point where stopping enhances rather than impedes a market's propensity to catapult people out of servility. (The colorful verb implicitly acknowledges that feeling free is not like feeling secure.)

Although Satz's criticism of noxious markets is measured and precise, there are hints of a more general critique. She says that because markets presuppose government regulations, "True laissez-faire is not even logically possible" (27). But she wisely does not rest much weight on trendy but implausible claims about how government regulations historically or logically precede the human propensity to make deals. To justify regulations, we need to focus on actual concerns, as Satz proceeds to do. It is a hallmark of liberal society that there is such a thing as minding one's own business. Yet, Satz says:

> The fact that laws and institutions underwrite market transactions also means that such transactions are, at least in principle, not private capitalist acts between consenting adults, as the libertarian philosopher Robert Nozick famously claimed, but instead a public concern of all citizens whether or not they directly participate in them. (16)

It feels awkward, though, to say that high on the list of things that are inherently public and therefore cannot count as minding one's own business is *literally minding one's own business*. Still, although we may wish to stop certain "capitalist actions between consenting adults," Satz never forgets that people we would protect (if we could) frequently are doing the best they can. Taking away someone's best option is one way of proving you care, but people who actually do care look for better alternatives. Satz also realizes that the noxiousness of some markets can be temporary. Infant industries and exploitative employers of first-generation immigrants may or may not foster rigidified class structure.

Satz also criticizes neoclassical economics for ignoring "the presence of massive political power that would undermine the market's optimizing effects" (61). She says, "The classical political economists viewed the economy not so much as a series of homogenous relations between individuals and the things they wanted, but as a system of heterogenous relationships between social classes" (60). Accordingly, the classical framework made it natural to treat the political process, democratic or not, as one that naturally would work in the interests of the kind of people who were adept at acquiring political power, and thus (to the extent that political redistribution is a zero-sum game) against the interests of the politically disenfranchised. A contemporary economist might dismiss Satz as uninformed, referring her to James Buchanan's and F. A. Hayek's Nobel Prize-winning analyses of the logic of political power. Classical economists like Smith and Marx, however, would applaud Satz for telling it like it is. For that matter, so would Buchanan and Hayek.

Neoclassical economists defend markets as vehicles for satisying our preferences, but Satz rejects this as circular, given that markets themselves can instill preferences (7, 49). Contra Satz on this point (and as Satz acknowledges in conversation), if a market enables us both to develop and to satisfy a taste for the music of Mozart, that is not circular.[1] We are indeed outcomes, not just makers, of our choices, and markets are crucibles in which values take shape. But what gives us bread and circuses also gives us Mozart. Not that every consumer likes Mozart, but enough do to bring his compositions into the realm of what is humanly possible. If there is a problem with economists, it is not that their reasoning is circular but that they do not acknowledge that we are the outcomes, not only the makers, of our choices. Just as Mozart's music is made possible by markets, so is the degradaton of a prostitute (143).

Smith was aware, and Satz is aware (ix), that capitalism makes formerly rigid class structures fluid, so that one day there would be nothing terribly unusual about poor and religious minority kids like Nozick and Satz growing up to become eminent professional intellectuals. In societies like ours, liberalism's victories against hereditary class privilege (and religious persecution) were so overwhelming that today we have no idea what a monster class privilege truly was, or that it did not go down without an apocalyptic fight (40, 42–48, 53–57).

---

[1] I thank Brandon Warmke for the thought.

Satz discusses the case of the Titanic. A ship, deemed unsinkable, sank without enough lifeboats for all of its passengers. The ship rationed lifeboat seats to passengers on the basis of their ticket class.[2] How is this an example of a good that should not be sold? Perhaps Satz's idea is that second-class citizenship should not be for sale. This fits with her suggestion that we might forbid ships from having a mix of passengers with unequal access to lifeboats: on a given ship, everyone sails with lifeboat access, or everyone sails without (89).

The world, of course, has a way of not cooperating with such ideals. Although Satz's remedy might typically induce everyone to think twice about risk, we also can imagine a case in which a dock is being engulfed by a volcano, and a would-be passenger's only hope is to board the Titanic. Unfortunately, as the ship's law-abiding captain explains, "We once had room for more passengers, but it now is illegal for us to carry more passengers than lifeboats. Please understand that these laws symbolize our concern for passenger safety." In this case, we took passenger safety out of the hands of freely transacting traders, rendering people helpless against the dangers of not transacting. Indeed, in the imagined case, the law recreated the very situation that motivated its passage. The ship itself has become the lifeboat, and the dock has become the sinking ship. So, while our permissions show what we stand for, so do our prohibitions.[3] There is no guarantee that prohibitions will have the result we intend, or even that they will come to symbolize concern rather than hubris. If the volcano example seems unrealistic, consider this question: what does prohibiting kidney sales symbolize to patients thus blocked from buying a life-preserving kidney?

Satz ponders how "treating and respecting people as equals" demands protecting them not only from exploitation but from their own mistakes as well (74). What is the connection, though, between being my brother's equal and being my brother's keeper?[4] I am not saying Satz got the wrong answer. Indeed, she shies away from any final answer; her aim is to warn against being overconfident in the ability of social policy to discern the difference in practice. Implicit in her critique of thinkers like Dworkin is an awareness that standing for equality is a lot easier than knowing which policies would launch a complex world toward an equality worth wanting.

---

[2] Were there fairer ways to allocate scarce seats than according to ticket class? One fair way would have been to draw lots. Indeed, we all share an intuition that arbitrariness (in the sense of randomness) is the essence of fairness in lotteries. The intuition is so powerful that one has to wonder how philosophers can swallow the premise that arbitrariness is the essence of *unfairness* when it comes to the natural "lottery."

[3] See Matt Zwolinski, "The Ethics of Price Gouging," *Business Ethics Quarterly*, XVIII, 3 (July 2008): 347–78.

[4] Suppose we believe, as many do, that ordinary people cannot be trusted to make choices for themselves, and should be overruled for their own good. Why not likewise say that when people vote, and make the wrong choice, we should overrule them for

We should note that law is not our only means for limiting markets. Our most powerful tools for resisting the intrusion of markets into the intimate aspects of our lives are markets themselves. Noxious markets in child labor are widespread only in poor countries. Satz supposes that minimum age (163) and minimum wage laws (209) can be used to make children unemployable.[5] Perhaps, but what we really want is not to raise minimum wages or ages but to raise living standards to a point where even the poor need not sell children to make ends meet.

We might say that "by far the most important effect" of free markets is not that they enable us to sell everything, but that they enable us not to. They make us rich enough that we can afford to say no. (Throughout the book, Satz leans toward whatever policies would be least paternalistic and most helpful. She would ask how to make being in school more salient and attractive. She might eschew blanket prohibitions of child labor, and instead favor laws against employing children during school hours. And so on.) It is within market societies that thick communities (Hutterite colonies, to give an extreme example), rich in uncommodified value, flourish. It is *only* within market societies that capitalism's critics flourish. Centuries ago, brave people gave their lives, liberty, and sacred honor to the project of inventing a society where intellectuals would have the right and the responsibility to take a hard look at what their country stands for. I think they would have regarded Satz's *Why Some Thing Should Not Be for Sale* as a sign that their sacrifice was not in vain. It is unmistakably the work of a mature philosopher and a free country.[6]

DAVID SCHMIDTZ

University of Arizona

---

their own good? (When we let people vote, are we letting them choose, or merely letting them vent?) See Jason Brennan, *The Ethics of Voting* (Princeton: University Press, 2011).

[5] See Guido Pincione and Fernando R. Tesón, "Self-Defeating Symbolism in Politics," this JOURNAL, XCVIII, 12 (December 2001): 636–52.

[6] I am grateful to Debra Satz for visiting the University of Arizona in March 2011 and for debating these issues with me before the Philosophy Department and my Ethics seminar. It is a tribute to Satz's grace and wisdom that, as my students and colleagues put it, "Everyone won that debate."

NEW BOOKS: ANTHOLOGIES

ABELL, CATHARINE and KATERINA BANTINAKI, eds.: *Philosophical Perspectives on Depiction.* New York: Oxford, 2010. xii, 241 p. Cloth: $75.00.

AGUILAR, JESÚS H. and ANDREI R. BUCKAREFF, eds.: *Causing Human Actions: New Perspectives on the Causal Theory of Action.* Cambridge: MIT, 2010. viii, 327 p. Cloth: $70.00, Paper: $35.00.

AHMED, ARIF, ed.: *Wittgenstein's* Philosophical Investigations*: A Critical Guide.* New York: Cambridge, 2010. ix, 248 p. Cloth: $85.00.

ANDERSON-GOLD, SHARON and PABLO MUCHNIK, eds.: *Kant's Anatomy of Evil.* New York: Cambridge, 2010. xi, 251 p. Cloth: $85.00.

ANDREOU, CHRISOULA and MARK D. WHITE, eds.: *The Thief of Time: Philosophical Essays on Procrastination.* New York: Oxford, 2010. xiv, 300 p. Cloth: $65.00.

ARCHARD, DAVID and DAVID BENATAR, eds.: *Procreation and Parenthood: The Ethics of Bearing and Rearing Children.* New York: Oxford, 2010. viii, 191 p. Cloth: $55.00.

AYALA, FRANCISCO J. and ROBERT ARP, eds.: *Contemporary Debates in Philosophy of Biology.* Malden, MA: Wiley-Blackwell, 2010. xii, 426 p. Cloth: $94.95, Paper: $41.95.

BAGGETT, DAVID, ed.: *Tennis and Philosophy: What the Racket is All About.* Lexington: Kentucky UP, 2010. viii, 283 p. Cloth: $35.00.

BAKER, MARK C. and STEWART GOETZ, eds.: *The Soul Hypothesis: Investigations into the Existence of the Soul.* Baltimore: Continuum, 2010. vii, 294 p. Paper: $19.95.

BEEBEE, HELEN, CHRISTOPHER HITCHCOCK, and PETER MENZIES, eds.: *The Oxford Handbook of Causation.* New York: Oxford, 2010. xv, 266 p. Cloth: $150.00.

BERGMANN, MICHAEL, MICHAEL J. MURRAY, and MICHAEL C. REA, eds.: *Divine Evil?: The Moral Character of the God of Abraham.* New York: Oxford, 2011. xi, 337 p. Cloth: $125.00.

BESSON, SAMANTHA and JOHN TASIOULAS, eds.: *The Philosophy of International Law.* New York: Oxford, 2010. xiv, 611 p. Cloth: $115.00, Paper: $40.00.

BETT, RICHARD, ed.: *The Cambridge Companion to Ancient Scepticism.* New York: Cambridge, 2010. xii, 380 p. Cloth: $95.00, Paper: $32.99.

BOUWEL, JEROEN VAN, ed.: *The Social Sciences and Democracy.* New York: Palgrave Macmillan, 2010. ix, 268 p. Cloth: $90.00.

BRICMONT, JEAN and JULIE FRANCK, eds.: *Chomsky Notebook.* New York: Columbia, 2010. vi, 351 p. Cloth: $89.50, Paper: $29.50.

BRIGHOUSE, HARRY and INGRID ROBEYNS, eds.: *Measuring Justice: Primary Goods and Capabilities.* New York: Cambridge, 2010. ix, 257 p. Cloth: $85.00, Paper: $29.99.

BRUCE, MICHAEL and ROBERT M. STEWART, eds.: *College Sex: Philosophers with Benefits.* Malden, MA: Wiley-Blackwell, 2010. xiv, 235 p. Paper: $19.95.

BRUNO-JOFRÉ, ROSA et al., eds.: *Democracy and the Intersection of Religion and Traditions: The Reading of John Dewey's Understanding of Democracy and Education.* Ithaca, NY: McGill-Queen's, 2010. 176 p. Cloth: $75.00, Paper: $29.95.

CAHN, STEVEN M., ROBERT B. TALISSE, and SCOTT F. AIKIN, eds.: *Thinking about Logic: Classic Essays.* Philadelphia: Westview, 2010. xii, 215 p. Paper: $24.00.

CAMPBELL, JOSEPH KEIM, MICHAEL O'ROURKE, and HARRY S. SILVERSTEIN, eds.: *Time and Identity.* Cambridge: MIT, 2010. vi, 330 p. Cloth: $35.00.

CLANTON, J. CALEB, ed.: *The Ethics of Citizenship: Liberal Democracy and Religious Convictions.* Waco, TX: Baylor UP, 2010. xix, 465 p. Cloth: $49.95.

COCHRAN, MOLLY, ed.: *The Cambridge Companion to Dewey.* New York: Cambridge, 2010. xviii, 356 p. Cloth: $85.00, Paper: $31.99.

COHEN, JONATHAN and MOHAN MATTHEN, eds.: *Color Ontology and Color Science.* Cambridge: MIT, 2010. xxiii, 419 p. Cloth: $90.00, Paper: $45.00.

COLMAN, FELICITY, ed.: *Film, Theory and Philosophy: The Key Thinkers.* Durham, UK: Acumen, 2010. xi, 404 p. Cloth: $120.00, Paper: $34.95.

COURTLAND, LEWIS and PAULA SMITHKA, eds.: *Doctor Who and Philosophy: Bigger on the Inside.* Chicago: Open Court, 2010. xviii, 429 p. Paper: $21.95.

DANCY, JONATHAN, ERNEST SOSA, and MATTHIAS STEUP, eds.: *A Companion to Epistemology: Second Edition.* Malden, MA: Wiley-Blackwell, 2010. xxiii, 799 p. Cloth: $204.95.

DAVIES, BRIAN, ed.: *God and Evil in the Theology of St. Thomas Aquinas.* Baltimore: Continuum, 2010. xviii, 205 p. Cloth: $24.95.

DAVIS, BRET W., ed.: *Martin Heidegger: Key Concepts.* Ithaca, NY: McGill-Queen's, 2010. xvi, 288 p. Cloth: $85.00, Paper: $24.95.

DAY, WILLIAM and VICTOR J. KREBS, eds.: *Seeing Wittgenstein Anew.* New York: Cambridge, 2010. xvi, 393 p. Cloth: $90.00, Paper: $27.99.

DE CARO, MARIO and DAVID MACARTHUR, eds.: *Naturalism and Normativity.* New York: Columbia, 2010. vi, 368 p. Cloth: $89.50, Paper: $29.50.

DENIS, LARA, ed.: *Kant's* Metaphysics of Morals*: A Critical Guide.* New York: Cambridge, 2010. xiii, 270 p. Cloth: $85.00.

DERANTY, JEAN-PHILIPPE, ed.: *Jacques Rancière: Key Concepts.* Durham, UK: Acumen, 2010. xii, 206 p. Cloth: $75.00, Paper: $24.95.

DIETZ, RICHARD and SEBASTIANO MORUZZI, eds.: *Cuts and Clouds: Vagueness, Its Nature, and Its Logic.* New York: Oxford, 2010. xi, 586 p. Cloth: $120.00.

DIXON, THOMAS, GEOFFREY CANTOR, and STEPHEN PUMFREY, eds.: *Science and Religion: New Historical Perspectives.* New York: Cambridge, 2010. xiv, 317 p. Cloth: $95.00.

DOMSKI, MARY and MICHAEL DICKSON, eds.: *Discourse on a New Method: Reinvigorating the Marriage of History and Philosophy of Science.* Chicago: Open Court, 2010. x, 852 p. Cloth: $89.95.

EDMONDS, DAVID and NIGEL WARBURTON, eds.: *Philosophy Bites.* New York: Oxford, 2010. xxiv, 244 p. Cloth: $15.95.

EGGLESTON, BEN, DALE E. MILLER, and DAVID WEINSTEIN, eds.: *John Stuart Mill and the Art of Life.* New York: Oxford, 2011. viii, 304 p. Cloth: $74.00.

ERNST, GERHARD and ANDREAS HÜTTEMANN, eds.: *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics.* New York: Cambridge, 2010. vi, 211 p. Cloth: $85.00.

FABER, ROLAND and ANDREA M. STEPHENSON, eds.: *Secrets of Becoming: Negotiating Whitehead, Deleuze, and Butler.* New York: Fordham, 2011. xx, 277 p. Cloth: $80.00, Paper: $28.00.

FAHY, THOMAS, ed.: *The Philosophy of Horror.* Lexington: Kentucky UP, 2010. vi, 259 p. Cloth: $35.00.

FEFERMAN, SOLOMAN, CHARLES PARSONS, and STEVEN SIMPSON, eds.: *Kurt Gödel: Essays for His Centennial.* New York: Cambridge, 2010. x, 373 p. Cloth: $90.00.

FLORIDI, LUCIANO, ed.: *The Cambridge Handbook of Information and Computer Ethics.* New York: Cambridge, 2010. xv, 327 p. Cloth: $95.00, Paper: $36.99.

FURTAK, RICK ANTHONY, ed.: *Kierkegaard's 'Concluding Unscientific Postscript': A Critical Guide.* New York: Cambridge, 2010. xiii, 258 p. Cloth: $85.00.

GALSTON, WILLIAM A. and PETER H. HOFFENBERG, eds.: *Poverty and Morality: Religious and Secular Perspectives.* New York: Cambridge, 2010. xvi, 312 p. Paper: $28.99.

GARDINER, STEPHEN M. et al., eds.: *Climate Ethics: Essential Readings.* New York: Oxford, 2010. xvi, 351 p. Cloth: $99.00, Paper: $35.00.

GARRARD, EVE and DAVID MCNAUGHTON, eds.: *Forgiveness.* Durham, UK: Acumen, 2010. xi, 132 p. Paper: $18.95.

GATTI, HILARY, ed.: *Essays on Giordano Bruno.* Princeton: University Press, 2010. xvii, 353 p. Paper: $75.00, Cloth: $35.00.

GENDLER, TAMAR SZABÓ, and JOHN HAWTHORNE, eds.: *Oxford Studies in Epistemology, Volume 3.* New York: Oxford, 2010. viii, 334 p. Cloth: $99.00, Paper: $35.00.

GERSON, LLOYD P., ed.: *The Cambridge History of Philosophy in Late Antiquity.* 2 vols. New York: Cambridge, 2011. xiii, 1284 p. Cloth: $240.00.

GILL, CHRISTOPHER, TIM WHITMARSH, and JOHN WILKINS, eds.: *Galen and the World of Knowledge.* New York: Cambridge, 2010. xvii, 327 p. Cloth: $99.00.

GILSON, CAITLIN SMITH, ed.: *The Metaphysical Presuppositions of Being-in-the-World: A Confrontation between St. Thomas Aquinas and Martin Heidegger.* Baltimore: Continuum, 2010. xiv, 219 p. Paper: $135.00.

GOLASH, DEIRDRE, ed.: *Freedom of Expression in a Diverse World.* New York: Springer, 2010. xxviii, 215 p. Cloth: $139.00.

GOLDIE, PETER, ed.: *The Oxford Handbook of Philosophy of Emotion.* New York: Oxford, 2010. xiv, 722 p. Cloth: $150.00.

GOLDMAN, ALVIN I. and DENNIS WHITCOMB, eds.: *Social Epistemology: Essential Readings.* New York: Oxford, 2011. ix, 357 p. Paper: $29.95.

GOTTHELF, ALLAN and JAMES G. LENNOX, eds.: *Metaethics, Egoism, and Virtue: Studies in Ayn Rand's Normative Theory.* Pittsburgh: University Press, 2010. xi, 200 p. Cloth: $30.00.

GRAMMONT, FRANCK, DOROTHÉE LEGRAND, and PIERRE LIVET, eds.: *Naturalizing Intention in Action.* Cambridge: MIT, 2010. viii, 350 p. Cloth: $35.00.

GUYER, PAUL, ed.: *The Cambridge Companion to Kant's* Critique of Pure Reason. New York: Cambridge, 2010. xiv, 461 p. Cloth: $90.00, Paper: $32.99.

HALE, BOB and AVIV HOFFMANN, eds.: *Modality: Metaphysics, Logic, and Epistemology.* New York: Oxford, 2010. vi, 363 p. Cloth: $85.00.

HAMBURG, G. M. and RANDALL A. POOLE, eds.: *A History of Russian Philosophy, 1830–1930: Faith, Reason, and the Defense of Human Dignity.* New York: Cambridge, 2010. ix, 423 p. Cloth: $120.00.

HARRISON, PETER, ed.: *The Cambridge Companion to Science and Religion.* New York: Cambridge, 2010. xi, 307 p. Cloth: $85.00, Paper: $24.99.

HAZLETT, ALLAN, ed.: *New Waves in Metaphysics.* New York: Palgrave MacMillan, 2010. x, 280 p. Paper: $35.00.

REVUE INTERNATIONALE DE PHILOSOPHIE

Direction: Michel Meyer
c / o Université Libre de Bruxelles, C.P. 188
50, avenue Franklin-Roosevelt / B-1050 Bruxelles (Belgique)
Tél. 32 2 650 25 61 / Fax 32 2 650 45 61
e-mail : tsalis@ulb.ac.be

1-2011 / volume 65 / n° 255

LE THEATRE. DE LA RENAISSANCE A AUJOURD'HUI. UNE APPROCHE PHILOSOPHIQUE

Michel MEYER
Présentation

Benjamin BENNETT
The Thinking Machine

Florence NAUGRETTE
Le mélange des genres dans le théâtre romantique français:
une dramaturgie du désordre historique

Michel MEYER
Du romantisme au réalisme dans le théâtre

Gérard NOIRIEL
L'institutionnalisation du théâtre français et ses effets
sur la définition légitime de la création

André HELBO, Catherine BOUKO, Elodie VERLINDEN
Théâtre et spectacle vivant. Mutations contemporaines

Prochains fascicules / Forthcoming
Stanley Cavell
Comte-Sponville
Autour de «Questionnement et Historicité»
William James

CAMBRIDGE

NEW *and* NOTEWORTHY TITLES *in* PHILOSOPHY

Ethics and Animals
An Introduction
LORI GRUEN
Cambridge Applied Ethics
$85 00 Hb 978-0-521-88899-8
$27 99 Pb 978-0-521-71773-1
250 pp

Kant: Observations on the Feeling of the Beautiful and Sublime and Other Writings
PATRICK FRIERSON *and* PAUL GUYER
Cambridge Texts in the History of Philosophy
$99 00 Hb 978-0-521-88412-9
$28.99 Pb 978-0-521-71113-5
396 pp

Philosophy and Temporality from Kant to Critical Theory
ESPEN HAMMER
Modern European Philosophy
$90 00 Hb 978-1-107-00500-6
270 pp

G. E. Moore: Early Philosophical Writings
THOMAS BALDWIN *and* CONSUELO PRETI
$110 00 Hb 978-0-521-19014-5
338 pp

Heidegger, Art, and Postmodernity
IAIN D THOMSON
$90 00 Hb 978-1-107-00150-3
$27 99 Pb 978-0-521-17249-3
266 pp.

Interpreting Heidegger
Critical Essays
DANIEL O DAHLSTROM
$90 00 Hb 978-0-521-76494-0
320 pp

Moral Dilemmas in Medieval Thought
From Gratian to Aquinas
M V DOUGHERTY
$90 00 Hb 978-1-107-00707-9
236 pp

Perception and Knowledge
A Phenomenological Account
WALTER HOPP
$90 00 Hb 978-1-107-00316-3
258 pp

Displacement by Development
Ethics, Rights and Responsibilities
PETER PENZ, JAY DRYDYK *and* PABLO S BOSE
$90 00 Hb 978-0-521-19882-0
$31 99 Pb 978-0-521-12464-5
358 pp

The Ethics of Global Climate Change
DENIS G ARNOLD
$90 00 Hb 978-1-107-00069-8
352 pp

Peirce and the Threat of Nominalism
PAUL FORSTER
$82 00 Hb 978-0-521-11899-6
272 pp

Socrates and Philosophy in the Dialogues of Plato
SANDRA PETERSON
$90 00 Hb 978-0-521-19061-9
310 pp

Max Horkheimer and the Foundations of the Frankfurt School
JOHN ABROMEIT
$95.00 Hb 978-1-107-00695-9
448 pp

Vygotsky in Perspective
RONALD MILLER
$110 00 Hb 978-1-107-00186-2
466 pp

Prices subject to change

Heidegger, Art, and Postmodernity
Displacement by Development

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 5
MAY 2011

page

227 *Primitive Normativity and Skepticism about Rules* — Hannah Ginsborg

COMMENTS AND CRITICISM

255 *Is Direct Reference Theory Incompatible with Physicalism?* — Mahrad Almotahari and Damien Rochford

REVIEW ESSAYS

269 Tyler Burge: *Origins of Objectivity* — John Campbell

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S. copyright law All rights reserved For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2011 by the Journal of Philosophy, Inc ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 5, MAY 2011

PRIMITIVE NORMATIVITY AND SKEPTICISM ABOUT RULES*

In his *Wittgenstein on Rules and Private Language*, Saul Kripke develops a skeptical argument against the possibility of meaning.[1] Suppose that all your previous uses of the word "plus" and of the "+" sign have involved numbers less than 57. You are now asked "what is 68 plus 57?" and you answer "125." But a skeptic proposes the hypothesis that by the word "plus," or the "+" sign, you previously meant not addition, but *quaddition*, where *x* quus *y* is the sum of *x* and *y* if *x* and *y* are less than 57, and otherwise 5. If you are to use the word "plus" as you used it in the past, the skeptic says, then, on the hypothesis that you meant quaddition, you ought to answer "5." Against your insistence that you know what you meant by "plus," the skeptic challenges you to cite some fact in virtue of which you meant addition. All your previous answers, he points out, were consistent with the hypothesis that you meant quaddition, so how can you justify your claim that you meant addition instead? The upshot of the skeptical

---

*This paper was originally written for the 2006 University of California, Riverside, Conference on Normativity and Universality from a Kantian Perspective. Subsequent versions were presented at Hofstra University; the University of Toronto; Harvard University; the University of Chicago's Wittgenstein Workshop; the 2007 University of Essex workshop on Normativity, Naturalism, and the Transcendental; the 2007 University of Kentucky conference on Normativity and the Mental; the Mind and Language Seminar at New York University; the Universidade Federale do Rio Grande do Sul in Porto Alegre; the Center for the Study of Mind and Language in Oslo; and seminars and discussion groups at the University of California, Berkeley, and at the Humboldt University in Berlin. I am grateful to participants on all those occasions for comments and discussion. I have benefitted in particular from exchanges with Ned Block, Paul Boghossian, Jason Bridges, Catherine Elgin, David Finkelstein, Barbara Herman, Paul Hoffman, Peter Hylton, Arpy Khatchirian, Sara Kisilevsky, Niko Kolodny, Christine Korsgaard, Jason Leddington, Beatrice Longuenesse, Andrews Reath, Karl Schafer, Barry Stroud, Daniel Sutherland, David Velleman, Jay Wallace, Daniel Warren, and Daniel Whiting. I would also like to thank the editors of this JOURNAL for very helpful comments

[1] Saul Kripke, *Wittgenstein on Rules and Private Language* (Cambridge: Harvard, 1982). Unless otherwise noted, page references are to this work.

0022-362X/11/0805/227-254

considerations is not merely the epistemological conclusion that you do not know what you meant, but the metaphysical conclusion that there is no fact about what you meant. And this conclusion generalizes to all supposed cases of meaning or rule-following, present as well as past. There can be no fact as to whether anyone means anything by any word, or is following any one rule rather than any other.

Kripke's argument proceeds mostly by elimination: he considers, and rules out, various proposals as to what the fact of meaning addition might consist in. Two in particular are worth highlighting. The one which Kripke discusses in more detail is that your meaning addition by the word "plus" is a fact about your dispositions with respect to that word. You meant addition in the past because you were disposed to give the sum rather than the quum in answer to questions using the word "plus." Kripke raises three objections to this proposal, turning respectively on the supposed finiteness of our dispositions (26–27), on the fact that we are sometimes disposed to make arithmetical mistakes (28–32), and, most importantly, on the proposal's apparent failure to account for what he calls the "normativity of meaning": that one is or was *disposed* to respond in a certain way on a given occasion cannot make it the case that one *ought* so to respond (23–24, 37).[2] The other, which Kripke treats much more briefly, is that your meaning addition by "plus" is a primitive or *sui generis* state (51). This proposal is antireductionist in spirit: the fact that you mean addition, on this proposal, cannot be reduced to facts about your behavior or dispositions, or to any nonintentional psychological facts about you. While Kripke acknowledges that this proposal may be "irrefutable," he rejects it as "desperate": it leaves, he says, the nature of the *sui generis* state "completely mysterious" (51).

Many of the responses to Kripke's skeptical puzzle have been aimed at defending one or the other of these two proposals. Kripke's attack on the dispositional view has been countered by invoking a notion of disposition on which dispositions are ascribed on the basis of behavior in ideal conditions, or subject to *ceteris paribus* clauses. We can say, for example, that salt has a disposition to dissolve in water even though it fails to dissolve in a saturated solution or when something interferes with the electric charges on the water molecules. This addresses both the finiteness objection and the objection regarding dispositions to make mistakes: we can say that I am disposed to give the sum even though that disposition is not actualized if I am given numbers that are too large for me to grasp or if my habitual carelessness interferes with my calculations.[3] And it has also been thought to address the objection regarding normativity, since the claim that a person is doing as she ought might seem to be grounded in the idea that she is responding as she would under ideal conditions.[4] In support of the proposal that meaning is a *sui generis* state, Kripke's critics have simply accused him of unargued reductionism.[5] Why should we suppose that the fact of someone's meaning something, or the fact of her following a rule, should be specifiable in purely naturalistic terms? These facts are, in a metaphor which John McDowell draws from Wittgenstein, "bedrock." We should not suppose that we can dig down below them to find more fundamental facts about verbal behavior and accompanying feelings on which the facts of meaning and rule-following rest.[6]

---

[2] Boghossian classifies both the second and the third objections under the heading of the "normativity of meaning"; he also regards the second objection as much the more important of the two. See Paul Boghossian, "The Rule-Following Considerations," *Mind*, XCVIII, 392 (October 1989): 507–49, at pp. 530–31. In referring to the "normativity" objection, I have in mind exclusively the third. On the distinction between the second and third objections, see for example Simon Blackburn, "The Individual Strikes Back," *Synthese*, LVIII, 3 (March 1984): 281–301, at p. 291; Crispin Wright, "Kripke's Account of the Argument against Private Language," this JOURNAL, LXXXI, 12 (December 1984): 759–78, at pp. 771–72; Scott Soames, "Skepticism about Meaning. Indeterminacy, Normativity, and the Rule-Following Paradox," *Canadian Journal of Philosophy*, Supplementary Volume, XXIII (1997) 211–49, at pp. 217–19.

But it is widely accepted that neither the dispositionalist nor the antireductionist account is fully satisfactory. Regarding the dispositionalist approach, it is not at all clear that the kinds of modifications that address the objections about finitude and mistakes also address the normativity problem. The idea that someone is responding or not responding as she would respond under ideal circumstances does not on its own license the idea that she is responding, or failing to respond, as she *ought*. We do not say of a sample of salt that dissolves or fails to dissolve in water that it is doing or not doing as it ought, but only that it is or is not manifesting its disposition to dissolve. So the dispositionalist account fails, on the face of it, to do justice to the intuition that a person's meaning something by a term, or her following a rule for its use, has normative implications.

---

[3] The response is developed in detail in Graeme Forbes, "Skepticism and Semantic Knowledge," *Proceedings of the Aristotelian Society*, n. s., LXXXIV (1983–1984): 223–37. See also Blackburn, *op. cit.*, pp 289–91; Wright, *op. cit.*, p. 772; and Jerry A. Fodor, *A Theory of Content and Other Essays* (Cambridge: MIT, 1990), pp. 94–95. I consider this line of response further in Hannah Ginsborg, "Inside and Outside Language: Stroud's Nonreductionism about Meaning," in Jason Bridges, Niko Kolodny, and Wai-hung Wong, eds., *The Possibility of Philosophical Understanding: Reflections on the Thought of Barry Stroud* (New York. Oxford, forthcoming).

[4] Warren Goldfarb, "Kripke on Wittgenstein on Rules," this JOURNAL, LXXXII, 9 (September 1985): 471–88, at pp. 477–78.

[5] See especially Colin McGinn, *Wittgenstein on Meaning* (Oxford: Blackwell, 1984), p. 150ff. Other antireductionist approaches include Boghossian, *op. cit.*; John McDowell, "Wittgenstein on Following a Rule," in *Mind, Value, and Reality* (Cambridge: Harvard, 1998), pp. 221–62; and Barry Stroud, "Mind, Meaning, and Practice," in *Meaning, Understanding, and Practice* (New York: Oxford, 2000), pp. 170–92.

[6] McDowell, *op. cit.*, §7, p. 241f.

While the antireductionist approach is a better candidate for accounting for the apparent normative implications of meaning states, it suffers from what is, in a sense, the converse difficulty. For it fails to account for the way in which what someone means, or what rule she is following, seems to determine not just what she *ought* to do, but what she in fact *will* do. Paul Boghossian raises a version of this general worry when he points out that the antireductionist view fails to account for the causal efficacy of meaning states. Another version of the worry is raised by Crispin Wright when he draws attention to the "disposition-like" connections to behavior which content-bearing psychological states exhibit. Such states, as Wright puts it, "resemble dispositions in the manner in which they have to answer to an indefinitely circumscribed range of behavioral manifestations."[7] At least on the face of it, as Wright points out, this quasi-dispositional character of meaning states is something for which the antireductionist has difficulty accounting.

My aim in this paper is to propose a solution to the skeptical puzzle which offers a middle way between these two approaches. This solution attempts to do justice to the way in which meaning and rule-following resemble dispositional states while still accommodating what Kripke calls the normativity of meaning. While my approach is partly reductionist, in that it aims to reduce facts about meaning to facts that are in a sense more primitive, it does not attempt a reduction of meaning to facts conceived purely naturalistically. "Bedrock" on this approach is located below the level of facts about meaning, but as we shall see it is still irreducibly normative, and hence it remains above the level of mere behavioral responses and their psychological concomitants. My solution centers on a notion which I call "primitive normativity" and which I take to be Kantian in origin. In the first of the four sections which follow, I introduce this notion in the context of the dialectic initiated by Kripke's skeptic. In section II, I say more about the notion of primitive normativity in its own right before going on, in section III, to show how it can be invoked to meet the skeptical challenge. In section IV I will address some objections and then go on to explain briefly why I take the view to be Kantian.

I

I want to begin by questioning an assumption which Kripke makes early in his development of the skeptical dialectic and which has not so far been challenged, to the best of my knowledge, by any of his critics. The assumption is implicit in a passage where Kripke describes the skeptical challenge as "taking two forms":

---

[7] Wright, "Critical Notice of Colin McGinn's *Wittgenstein on Meaning*," *Mind*, XCVIII, 390 (April 1989): 289–305, at p. 293.

> First, [the skeptic] questions whether there is any *fact* that I meant plus, not quus, that will answer his sceptical challenge. Second, he questions whether I have any reason to be so confident that now I should answer '125' rather than '5'. The two forms of the challenge are related. I am confident that I should answer '125' because I am confident that this answer also accords with what I *meant*. (11)

Now, as the development of the dialectic makes clear, Kripke's main concern is with the first of these challenges: to show that there is a fact that you meant "plus."[8] The relevance of the second challenge is that it imposes a constraint on any satisfactory answer to the first challenge. If any fact counts as the fact of your meaning addition, it must serve to justify your confidence that you ought to say "125" rather than "5."

What I want to question here is the way in which Kripke takes these two challenges to be related. Kripke assumes that the first challenge must be met as a prior condition of responding to the second: in order to claim legitimately that you ought to say "125," you need first to establish that you previously meant addition. This is because he assumes that your claim about what you ought to say must rest on a claim about what you meant. As he puts it in the passage just quoted, "I am confident that I should answer '125' *because* I am confident that this answer also accords with what I meant" (11, my emphasis). The skeptic can thus challenge your confidence about the former of these by challenging your confidence about the latter. On this assumption, a response to the skeptic must proceed by first showing, against the skeptic's first challenge, that you did indeed mean addition. Only then are you in a position to claim, in response to the skeptic's second challenge, that you ought to say "125" if you are to accord with your previous usage. But I want to deny that this is the order in which the two challenges must be addressed. I want to propose that you can legitimately reply to the skeptic that you ought to say "125" independently of any assumption about what you, or indeed anyone, meant previously by "plus." That means that you can dismiss the second challenge by pointing out that it is unmotivated: you can maintain your previous confidence that "125" is the appropriate answer regardless of whether or not you can answer the first challenge.[9] Now, as we saw, it is the first challenge which is most central to Kripke's skeptic, so this proposal on its own does not yet amount to a reply to the skeptical puzzle. But I shall argue in section III that if you can respond to the skeptic's second challenge in the way I have suggested, that puts you in a position to show that you meant addition rather then quaddition and, hence, to answer the skeptic's first challenge as well.

---

[8] Although Kripke refers in this passage to "two forms" of the skeptical challenge, he has the skeptic present them as two distinct, albeit related, challenges, and this is how I will treat them in what follows

It is important to be clear that the skeptic's second challenge concerns not how you ought to respond *simpliciter*, but how you ought to respond in light of your previous use of the term "plus." You and the skeptic agree that you previously responded to "2 plus 3" with "5," to "7 plus 5" with "12," and so on. Your disagreement concerns whether "125" is what you ought to say relative to that past history of use. The question, as Kripke puts it, is "whether my present usage agrees with my past usage" (12): it is not whether "125" is now the answer which should be given to "68 + 57," but rather whether "125" is what "[I should say] to accord with my previous usage" (12). Kripke treats interchangeably the idea that "125" is what you ought to say if you are to accord with your past *usage*, and the idea that this is what you ought to say if you are to accord with your past *meaning* or with your past *intentions*. He glosses the skeptic's question of "whether my present usage agrees with my past usage" as the question "whether I am presently conforming to my previous linguistic intentions" (12; Kripke's emphases removed); the question is, "How do I know that '68 plus 57', as I *meant* 'plus' in the *past*, should denote 125?" (12). This is because he takes for granted that accordance with how you used the term in the past can be understood only as accordance with the meaning with which you used the term, or as accordance with the rule which you adopted for its use. But I am denying that the idea of conformity to past usage depends on the idea of conformity to past meaning. While I accept that the "ought" in question is conditional on the circumstances in which you used the word "plus" in the past, I reject the assumption—which is of a piece with Kripke's assumption about how the skeptic's two challenges are related—that the "ought" has to be conditional on your past meaning or past intentions, or on a rule which you previously had in mind for the use of the term. I maintain that there is a sense in which you ought to say "125," given

[9] I am assuming here, with Kripke, that the past and present meanings of all signs except those referring to addition can be taken for granted. When you say "125" you can be understood either as saying the name of the number 125, or as asserting that the answer to the question you were asked is "125." *Pace* Soames, *op cit.*, p. 217, you do not have to be understood as asserting that 68 plus 57 is 125.

the finite list of your previous uses, independent of what meaning, if any, those uses expressed.

II

The sense of "ought" I am invoking here expresses what I am going to call "primitive normativity": very roughly, normativity which does not depend on conformity to an antecedently recognized rule. To bring this notion into focus I want to begin by considering a simpler analogue of Kripke's skeptical challenge, based on the example Wittgenstein gives at §185 of the *Philosophical Investigations*.[10] Imagine a child who is familiar with the numerals and able to recite them well into the hundreds, and who has just now learned to count by twos, that is, to recite numerals in the sequence 2, 4, 6, 8, and so on. Suppose that on one occasion she recites the sequence up to "40," and then goes on, as we expect, with "42." Moreover, she does so unhesitatingly, with an apparent assurance that this is the appropriate continuation. Now we stop her. "Why did you say '42'? Shouldn't you have said '43' instead?"

How we imagine her replying will depend on how we imagine that she learned her new skill. We might suppose that before learning to count by twos she was familiar not only with the numerals, but also with the words "plus" and "addition," and relatedly, that she was able to give answers to simple addition problems. In that case she might have been taught to count by twos by being given successive addition problems: she was asked to add two to 2, then to 4, then to 6, and so on. A child who learned counting by twos in this way could explain that she had said "42" because she had been adding two and because two added to 40 makes 42. Although she might not use words like "rule" or "justification," this would amount to a justification of her response in terms of its according with a rule which she had adopted antecedently. She would be citing a rule which she had been following up to this point (the "add-two" rule) and claiming that, in this particular case, the number she had given was an application of that rule.

But we could also imagine the child having learned to count by twos without receiving any specific instructions, in the same way that she learned to count by ones, that is, to recite the series of natural numbers. A child does not learn to count by following instructions like "Add one to the previous number," but rather by following the example given by other people and responding appropriately to their encouragement or correction. She learns initially by rote memorization, first of the numbers up to 19 and then of the decades (20, 30, and so on), but at a certain point she becomes able to recite sequences without relying exclusively on memory, at which point it becomes clear that she has acquired a capacity to count on her own.[11] We could imagine a child learning to count by twos in much the same way, by listening to other people reciting "2, 4, 6, 8..." and following their example. Such a child could go on confidently with "42" after "40," even if she had not heard this sequence before, or did not remember it, without being able to answer or even understand the question, "What is 40 plus 2?"

---

[10] Ludwig Wittgenstein, *Philosophical Investigations* (New York: Macmillan, 1953).

Now, it is possible that a child who had learned to count by twos in this way could come up with an explanation of why she said "42" after "40." In particular, she might have arrived on her own at a conception of what she was doing at each step of the process, and she might now be able to articulate that conception by saying, for example, that each number she said was "two more" than the one before, and that 42 came next because it was two more than 40. But it is at least equally as likely that the child would be unable to explain or justify her having said "42," not just because she lacked the appropriate vocabulary, but because she lacked any conception of what her saying "42" after "40" had in common with her having said "40" after "38." And yet this does not seem to rule out her reacting with surprise and puzzlement to the suggestion that she should have said "43" instead. Rather, it seems plausible to imagine her insisting, with no less conviction than a child who was able to cite the add-two rule, that "42" was the right thing to say after "40": that it "came next" in the series, or "belonged" after 40, or "fit" what she had been doing previously.

Suppose that the child does react in this way. On the face of it, there is a normative claim implicit in her reaction: she is claiming that "42" is appropriate, or what she ought to say, given what she has said previously. But this claim does not appear to depend on any prior claim to the effect that her response conforms to a rule. If we take this appearance at face value, then the child has made a claim to what I call primitive normativity. The utterance, from her point of view, is not appropriate to the context in virtue of its conforming to a general rule which the context imposed on her, for example, the add-two rule. Rather, she takes it to be appropriate to the context *simpliciter*, in a way which does not depend for its coherence on the idea of an antecedently applicable rule to which it conforms. This is not to deny that the normativity depends on any facts about the context, since the appropriateness of "42" depends on her having recited that particular sequence of number words. But it *is* to deny that her claim to the appropriateness of "42" depends on her recognition of a rule imposed by the context in virtue of the relevant facts, or *a fortiori* on her recognition of "42" as a correct application of the rule.[12]

---

[11] For details, see Thijs Pollmann, "Some Principles Involved in the Acquisition of the Number Words," *Language Acquisition*, xi, 1 (2003) 1–31. Learning to recite the sequence of number words is distinct from, and apparently to some extent precedes, the capacity to understand what they mean I overlook this complication in what follows.

I began with the child counting by twos both because of the familiarity of Wittgenstein's example on which it is based and because Wittgenstein's example presumably served as the prototype for Kripke's quaddition case. However, the phenomenon I am illustrating is not restricted to numerical examples, but pervades concept acquisition and language learning more generally. Imagine a child who is not yet obviously in command of color concepts—for example, who has not yet mastered use of the word "green"—but who, following an adult's example, is successfully sorting a collection of variously shaped and colored objects so that all the green ones go in one particular box. As she puts each green object in the designated box, it is plausible that she does so with a sense that this is the appropriate thing to do. She takes it that the green spoon "belongs" in the box containing the previously sorted green things and that the blue spoon does not, just as the child in the previous example takes 42 and not 43 to "belong" after 40 in the series of numerals. But her sense of the appropriateness of what she is doing does not, at least on the face of it, depend on her taking what she is doing to accord with a rule which she was following, for example, the rule that she is to put all the green things in the same box. For her grasp of such a rule would presuppose that she already possesses the concept *green*. The same holds if we suppose that her sorting behavior is linguistic, that is, if instead of putting each green object in a box, she says "green" when she is shown it or points to it when the teacher says "green." A child learning to use linguistic expressions in this way is not conscious of a rule which specifies, for example, that when a teacher says "green" she is to point to a green object. But this does not prevent her taking her response to the teacher's utterance to be appropriate, that is, to "fit" what the teacher has just said. She can hear the teacher's utterance as, so to speak, calling for her to point to the green object, without taking her response to conform to the meaning of, or to a rule for the use of, "green."

---

[12] For a further illustration of primitive normativity in a numerical context, consider children who learn arithmetic using Cuisenaire rods, which are wooden rods of different lengths proportional to the numbers one through ten. A child who has successively laid out the two-, four-, six-, and eight-unit rods might be confident that it is the ten-unit rod which should come next, and that another child who wants to put down the nine-unit rod is going on "wrong," without yet grasping that each rod in the pattern is two units longer than the previous one.

It will be useful here to compare the child's learning to recite series of numerals or sort objects by color with an animal's being trained to produce discriminative responses. Consider, say, a parrot being taught to say "green" when shown a green thing, or to peck at a green thing when the trainer says "green." This kind of training typically results in the parrot acquiring a disposition to say "green" when, and only when, a green thing is presented to it. In the case of the parrot we can distinguish two factors which go into its acquisition of the disposition. One factor is the training it receives. The characteristics of this training are responsible for the parrot's being disposed to produce the particular responses it does, for example, to say "green" rather than "grün" when shown a green thing, or, when the trainer says "green" in the presence of something green, to peck at, rather than step on, the green thing. The other factor is the parrot's innate disposition to "go on" from its training in a way which tracks the presence of greenness, as opposed to any of the other features common to the objects in the training sample. Suppose that the parrot's initial training was with a sample of objects which were not only green, but also opaque and inedible. A typical parrot trained to associate "green" with the objects in this sample will go on to say "green" not only when shown objects which are green, opaque, and inedible, but also when shown a translucent green cup or a green apple. The fact that the parrot's training results in a disposition to associate the "green" response with all and only green things, as opposed to things which which are green, opaque, and inedible, is not a consequence of the training itself, but rather of what we might think of as an innate second-order disposition to respond to its training in one way rather than another.

Now, I am assuming that these two factors also operate in the case of the child who learns to sort things by color or to count by twos. It is worth emphasizing the role of the second in particular. That the child recites series of numbers at all, or that she begins with "2, 4, 6, 8..." in response to the words "add two" as opposed to some other expression, is a consequence of the training she has received. But her being disposed to go on after "40" with "42" rather than "43" is a reflection not of how she has been trained, but of her natural tendency to go on in a way which tracks the series of even numbers. We can conceive of an abnormal child, like the pupil in Wittgenstein's example at *Investigations* §185, to whom it "comes naturally" to respond to the same training by going on with "43" rather than "42." Similarly, we can conceive of an abnormal child who responds to the standard training by using "green" as we do, except that she withholds it from green spoons and applies it to spoons which are blue. No amount of training, however, short of giving explicit instructions about what to say after "40," or how to apply "green" to spoons, could bring a normal human child to recite numbers or use the word "green" in these nonstandard ways. As in the case of animals, it is a fact about the nature of our species—a nature which, at least as regards our ability to sort by color, many animals share—that we acquire the first-order counting and sorting dispositions that we do.

It is part of my proposal that a child's continuing the series with "42" or applying the word "green" to a green spoon can be explained in the same naturalistic way that we explain the parallel behavior in the case of the parrot. But, I suggest, the situation of the child differs from that of the parrot in that the former takes herself, in continuing the series with "42" or saying "green" when shown the green spoon, to be responding appropriately to her circumstances in the primitive sense of "appropriate" which I have described. This consciousness of appropriateness does not explain why she goes on with "42" or applies "green" to the spoon, since, as I have just indicated, this is sufficiently accounted for in terms of her natural second-order dispositions to respond to training one way rather than another, and this is the kind of disposition which also could be shared by an animal which lacked any consciousness of normativity. On the contrary, it is because her natural dispositions lead her to go on with "42" rather than "43" that it is "42" rather than "43" which she takes to be the appropriate continuation. But it still makes a difference to her behavior, which we can put metaphorically by saying that, unlike the parrot, she does not respond "blindly" to her circumstances. Even though she does not say "42" as a result of having grasped the add-two rule, nor *a fortiori* of having "seen" that 40 plus two is 42, she nonetheless "sees" her utterance of "42" as appropriate to, or fitting, her circumstances.

My proposal represents a middle ground between two alternative, contrasting interpretations of the phenomena to which I have been calling attention. One interpretation would deny that the child adopts a genuinely normative attitude to her own behavior. What appears to be a sense of the appropriateness of what she is doing is in reality a complex of feelings without normative content: for example, pleasure at the prospect of an adult's approval. On this interpretation there is no qualitative difference between the child's response to her circumstances and the parrot's, only a difference in the degree of psychological refinement they manifest. The contrasting alternative would allow that the child, in saying "42," genuinely takes her utterance to be appropriate, but it would deny that this claim is, as I have put it, "primitive." On this alternative, the child says "42" after "40" because she recognizes, although without being able to put that recognition into words, that she has been adding two and that 40 plus two is 42. Her sense of the appropriateness of what she is saying thus derives from her recognition that it fits the rule she was following: a rule which she grasps, even though she is unable to articulate it.

I do not think that either of these alternatives can be ruled out definitively. My primary reason for preferring the middle ground lies in the solution it promises to Kripke's skeptical problem. The dispositionalist solution available on the first alternative and the antireductionist solution available on the second are both unsatisfactory, and the main advantage I claim for my proposal is that it offers a middle way between them. But the respective shortcomings of these interpretations can be characterized without explicit reference to the skeptical problem, in terms of their inability to account for the apparent role of activities like counting or sorting by color in the child's acquisition of concepts and rules. If, as on the second interpretation, the child needs to grasp the add-two rule or the concept *green* as a prior condition of being able to count by twos, or to distinguish the green from the nongreen objects, then her becoming competent in these activities could not be a way of coming to grasp the add-two rule or to acquire the concept *green*. In getting better at these activities, she would be getting better at manifesting her grasp of the relevant rule or concept, but for an explanation of how she had come to grasp the rule or concept in the first place, we would need to look elsewhere. But if, as on the first interpretation, we suppose that there is nothing more to the child's learning to sort or count than her acquisition of a parrot-like disposition to respond discriminatively to green things or to recite a series of numerals which conform to the add-two rule, then it is equally hard to see how her activity could contribute to a grasp of the corresponding rule or concept. For unless we simply identify grasp of the concept *green* with a capacity to discriminate green things, there is no reason to suppose that the acquisition of that capacity should put the child in a position to grasp the concept. What seems to be needed, if her becoming competent in the activity of sorting green things is to amount to her "catching on" to what green things have in common, is that, in acquiring that competence, she comes to see the green things as in some sense "belonging" together. And that idea of "belonging" is just what the element of normativity in my account tries to accommodate.

There are, indeed, other ways one might try to occupy a middle ground between the two alternatives I have described. One approach would allow, as on my proposal, that the child takes "42" to be appropriate in a way which does not rest on conscious recognition of the add-two rule, but it would ascribe her sense of appropriateness to an implicit grasp of that rule. Here, however, we need to be clear about what it is to grasp or follow a rule implicitly. Implicit rule-following is typically conceived as the operation of a "subpersonal" mechanism, where the rule is viewed as a representation to which the subject's cognitive system, but not the subject herself, has access. On that conception, I would grant that the child may be engaged in implicit rule-following and, more specifically, that there might be a mental representation corresponding to the add-two rule somewhere in her cognitive system, but I would deny that her taking "42" to be appropriate could be based on her recognition that it accords with that representation. For that would require that the "rule," that is, the representation, be accessible to her at the personal level. The hypothesis of implicit rule-following, on this subpersonal understanding, could help account for her actually saying "42"—perhaps, in both her case and the parrot's, training by reinforcement works precisely by building up the subpersonal psychological mechanisms in which implicit rule-following consists—but the hypothesized rule could not underwrite her claim to its appropriateness.[13]

Another possibility would be to account for the child's normative attitude in terms of the notion of sameness or similarity. The child takes "42" to be appropriate because she takes it that, in saying "42," she is doing the same thing as or something similar to what she was doing before.[14] Now, it might be objected against this approach that we cannot make sense of the notions of similarity or sameness without supposing some specific respect in which the items are similar, so that the child's claim to be doing the same thing is empty unless she conceives of herself, for example, as adding the same number. But I do not take this objection to be decisive in the present context, since if we are allowing ascriptions of primitive normativity, which do not depend on the assumption of some respect in which a given performance is appropriate, then there seems to be no ground for excluding a parallel notion of "primitive similarity," which can be ascribed without reference to a respect in which the items are similar. There is, however, a further difficulty, which I think is decisive: namely, that the child's recognition of similarity is not sufficient to account for her taking herself to be going on appropriately. She must not merely take herself to be going on the same way; she must also take it that going on the same way is the appropriate thing to do in the context, which is to say that she must grasp a rule with a content like *go on the same way* or *do the same thing you were doing before*. We are thus left with the problem of how to account for her grasp of this rule, and a related difficulty in addressing Kripke's skeptical problem. For if her claim to the appropriateness of "42" is based on the claim that it accords with the rule *do the same as before*, then it is open to the skeptic to claim that by "same" she meant *quame*, where doing the *quame* requires saying "43" instead.[15]

---

[13] The proposal I offer in section III might be said to ascribe "implicit" rule-following to the child, but in a different sense from the subpersonal sense described here.

[14] This is suggested by David Lewis: "adding means going on in the same way as before when the numbers get big, whereas quadding means doing something different." Lewis, "New Work for a Theory of Universals," *Australasian Journal of Philosophy*, LXI, 4 (1983) 343–77, at p. 376 For Lewis, this is not only the "naive" solution to Kripke's rule-following puzzle, but also the "correct" one.

I have introduced the notion of primitive normativity with examples of children, because the case for ascribing consciousness of primitive normativity in the performance of a given activity is clearest for people who have not yet mastered the relevant rule or concept. But I take it to be no less applicable in the case of adults. That an adult counting by twos conceives of herself as following the add-two rule, and takes "42" to accord with that rule, does not exclude her taking "42" to fit the preceding series *simpliciter*, in a way which does not depend on the assumption that she was following the add-two rule rather than a quadd-like variant.[16] Moreover, I take it not just that we in fact do make claims to primitive normativity, but that we are entitled to do so. I am committed not merely to the truth of the anthropological claim that human beings are disposed to take "42" to be appropriate in the circumstances described, but to the legitimacy of the normative attitude it ascribes. While this may look like a strong theoretical commitment, it is, as I see it, merely the articulation of a pretheoretical intuition which we all share. We would have no hesitation in telling a child who has counted by twos up to 40 and now wants to know "what comes next" or "what she ought to say now" that she should say "42," regardless of whether it has been specified that she is adding two each time. And, outside of a philosophical context, it would not occur to us to question one another's entitlement to answer the child's question in that way. The same pretheoretical intuition underlying our judgment that "42" is appropriate to the preceding series, independently of any assumptions about which rule the series instantiates, underwrites the claim that such judgments are legitimate. It is on this kind of intuition that my proposed response to the skeptic's second challenge relies. Regardless of what you meant by "plus" in the past, I propose, your saying "125" in response to the "68 plus 57" query is appropriate in light of your previous responses to "plus" questions in the same way that the child's saying "42" is appropriate in light of the preceding sequence of numerals.

---

[15] Cf. Kripke on "quimilar," p. 59n45. This kind of challenge is ineffective against my own proposal, since the child's claim to be going on appropriately does not rest on the assumption that she antecedently gave herself a rule to "go on appropriately." What she previously meant by "appropriate"—in particular, whether she meant *appropriate* or *quappropriate*—is thus irrelevant to her present claim to be going on appropriately.

[16] In fact, as I shall argue in section III, our grasp of the rules in terms of which we justify responses in particular cases is not only compatible with, but depends on our being able to take responses like "42" to be appropriate in the primitive sense.

III

I have argued so far for a response to the skeptic's second challenge which does not depend on an answer to his first, and more important, challenge. Now I want to show that the notion of primitive normativity can be used to answer the first challenge as well. My proposal for answering that challenge can be seen as a modification of the dispositional view, so I want to begin by looking at what I take to be the most serious objection to that view. Recall that Kripke raised three objections to the dispositional view, turning respectively on the supposed finiteness of dispositions, on the fact that people can be disposed to make mistakes, and, most importantly, on the dispositional view's supposed failure to accommodate the normativity of meaning and rules. For the purposes of this paper, I assume that the first two objections can be addressed.[17] This still leaves, however, the normativity objection: namely, that if one's meaning something by a term is identified with a disposition to use that term in a certain way, then we cannot account for the normative relation between what one means by the term and how one uses it. Kripke summarizes the objection as follows:

> Suppose I do mean addition by '+'. What is the relation of this supposition to the question how I will respond to the problem '68 + 57'? The dispositionalist gives a *descriptive* account of this relation: if '+' meant addition, then I will answer '125'. But this is not the proper account of the relation, which is *normative*, not descriptive. The point is *not* that, if I meant addition by '+', I *will* answer '125', but that, if I intend to accord with my past meaning of '+', I *should* answer '125'....The relation of meaning and intention to future action is *normative*, not *descriptive*. (37)

This objection has come under a lot of criticism, in particular for its assumption of what has come to be called the "normativity of meaning."[18] But I think that much of the criticism can be answered if the normative relation between meaning and use is understood properly. Before offering a positive account of that relation, I make two negative points. First, the relevant notion of normativity is not that associated with rationality. The claim that one "should" or "ought to" do something is often understood as implying that one has reason or, more strongly, conclusive reason to do it. But we can also speak of how one "ought" to use a term in a thinner sense which correlates with talk of how the term "ought to be used" and which also can be captured in terms of correct, as opposed to incorrect, use. How one ought to use a term in this sense does not depend on how one has reason to use it, since we can make perfectly good sense of someone having reason to use a term incorrectly. But unless we are reserving the term "normative" for the demands of rationality, this is no obstacle to describing the relevant "ought" as having normative content.[19] Second, the meaning with which one uses a term at one time imposes no normative requirements, even in this thinner sense, regarding subsequent use. Kripke's reference to "future action" notwithstanding, the fact that you now mean addition by "plus" does not make it the case that you should say, or that it will be correct for you to say, "125" in the future. It indeed does make it the case that if you do not say "125" you will not be conforming to what you now mean by "addition." But unless we make the further assumption that it is correct for you to mean the same by "plus" in the future as you do now, then nothing follows regarding what you should say in the future.[20] The same is true in the case of rules more generally. That you have now adopted a rule to behave in a certain way makes it the case that if you do not behave that way in the future you will not be conforming to the rule, but it does not follow that you ought, or that it is correct for you, to behave that way in the future.

---

[17] I defend this assumption in Ginsborg, *op. cit.*

---

[18] Challenges to the normativity of meaning and, relatedly, mental content, can be found in George Wilson, "Kripke on Wittgenstein on Normativity," *Midwest Studies in Philosophy*, XIX, 1 (1994): 366–90, David Papineau, "Normativity and Judgement," *Aristotelian Society Supplementary Volume*, LXXIII (1999): 16–43; Åsa Wikforss, "Semantic Normativity," *Philosophical Studies*, CII, 2 (January 2001): 203–26; Boghossian, "The Normativity of Content," *Philosophical Issues*, XIII, 1 (October 2003): 31–45; Anandi Hattiangadi, *Oughts and Thoughts* (New York: Oxford, 2007), and Kathrin Glüer and Wikforss, "Against Content Normativity," *Mind*, CXVIII, 469 (January 2009): 31–70.

[19] Here I agree with a point made by Gideon Rosen at pp. 620–21 of "Brandom on Modality, Normativity and Intentionality," *Philosophy and Phenomenological Research*, LXIII, 3 (November 2001): 611–23.

[20] Millar argues that meaning involves a commitment which is (as John Broome puts it) "wide-scope": your meaning what the community means by a term makes it the case that you ought either to use the term in accordance with the communal practice for the use of the term, or to abandon the practice. See Alan Millar, "The Normativity of Meaning," in Anthony O'Hear, ed., *Logic, Thought and Language* (New York: Cambridge, 2002), pp. 57–73. But it is not clear what it would be for you to abandon the practice, other than for you to stop using the term in accordance with the practice, and if that is all it takes, then the commitment is not a substantive one, since you can discharge it simply by using the word differently.

This second point raises the question of whether there is anything normative about the notion of a rule. If your adopting a rule creates no normative commitments regarding your future behavior, then it might be concluded that there is nothing more to the idea of a rule than that of a description which your behavior in a given instance might or might not satisfy. But this conclusion mislocates the normative element in the notion of a rule. The notion of a rule has normative content not because your adoption of a rule at one time makes it correct rather than incorrect to behave a certain way in the future, but rather because your adoption of a rule licenses or makes intelligible the application of the concepts of correctness and incorrectness to your behavior in the first place.[21] That someone has adopted or is following a certain rule creates a context in which we can sort her behavior not merely into behavior that does or does not meet a certain description, but into behavior that is correct or incorrect. We can put this by saying that a rule is something which stands in a normative relation to behavior, such that certain instances of behavior count as contravening or violating the rule and others as conforming to it (in a normative sense which contrasts with contravening, and which we might convey with terms like "comply with," "obey," or "respect"). But it is important to keep in mind that this normative relation to a rule does not bring it about that instances of your behavior are correct or incorrect except in the sense that it enables them to *count* as correct or incorrect, where this contrasts with their merely satisfying or not satisfying the corresponding description. The same holds in the case of meaning. Meaning stands in a normative relation to use not because your meaning something by a term now creates a commitment with respect to future use, but because it makes it intelligible to characterize your uses of the term as correct or incorrect. Regardless of whether or not "plus" has a meaning for you or anyone else, we can characterize your responses to "plus" questions by saying that they do or do not give the sum. But it is only on the assumption that "plus" means addition that we can describe your responses as correct or mistaken. The normativity of rules and meanings thus understood is extremely minimal, but it is enough to exclude the possibility of simply replacing talk of rules and meanings with talk of dispositions. For the notion of a disposition as such does not license any normative characterization of the behavior to which one is disposed. If you are disposed to use "plus" in such a way as to respond to "plus" questions with the sum, we can characterize your uses of "plus" not only as either giving or not giving the sum, but also as actualizing or not actualizing your disposition. But we still cannot, without further assumptions, characterize you either as responding correctly or as making a mistake.[22]

---

[21] I am using "correct" here in a sense which contrasts with "incorrect," that is, on which the denial of correctness amounts to the ascription of incorrectness. The idea of primitive normativity presented in the previous section also might be formulated in terms of "correctness," as the idea that we can take responses to be correct in a given context without the recognition of a rule in virtue of which they qualify as correct. But that would be to think of correctness as contrasting not with incorrectness, but with a broader notion of not being correct, where behavior that is not correct might include not only behavior which we might later classify as mistaken, but also behavior which is not subject to normative evaluation at all. If we understand the notion of correctness in that way, then we can regard your behavior as correct independently of the assumption that you have adopted a rule. Rules then can be regarded as having normative significance because their adoption makes intelligible a characterization of your behavior as incorrect or mistaken.

With this understanding of the normativity objection in place, I now want to suggest a way in which the dispositional view can be modified to meet it. This is to follow the standard dispositional line of identifying the state of meaning addition by "plus" with that of being disposed to give the sum in answer to queries using the word "plus," but to add the proviso that in actualizing that disposition, one takes oneself to be responding as one ought in the primitive sense. On this suggestion, you mean addition by "plus" if you are disposed to respond to a query about (say) "68 plus 57" with "125," where, in responding that way, you take that response to be primitively appropriate in light of your previous uses of "plus." The normative proviso builds into your disposition the feature that every response you are disposed to give involves a claim to its own appropriateness to the context in which you give it. Thus, your disposition is not just to say "125" in answer to "68 plus 57," "126" in answer to "68 plus 58," and so on; it is also, in each case, to take what you are saying to be the appropriate response to the query. You are disposed not only to respond with a number which is in fact the sum, but to consider that particular response appropriate.

---

[22] The account I have given makes no mention of the idea that meaning guides, justifies, or instructs us in the use of the word, although, as emphasized in Stroud, *op. cit.*, and in Bridges, "Rule-Following Skepticism, Properly So Called," in James Conant and Andrea Kern, eds., *Varieties of Skepticism: Essays after Kant, Wittgenstein, and Cavell* (Berlin: De Gruyter, forthcoming), this is part of Kripke's own understanding of the relation between meaning and use. Bridges takes the idea to be essential to the normativity objection, rejecting the account I have offered on the grounds that it fails to rule out the antireductionist approach (§2). But since I do not take the objection to be aimed against antireductionism, I think that it can and should be articulated independently of the idea that meaning justifies use. Readers who disagree may take my account as a sympathetic reconstruction rather than an accurate rendering of Kripke's view.

If your meaning addition is understood in this way, then, I suggest, it allows us to make sense of a given response being correct or incorrect. For we can intelligibly describe your various possible responses on any occasion not only as cases of your giving or not giving the sum, or of your actualizing or not actualizing your disposition to give the sum, but as your doing or not doing what you are disposed to regard as appropriate to the context. And the idea of your doing or not doing what you are disposed to regard as appropriate gives what is needed for you to count as responding correctly or making a mistake. If, in responding to the "plus" question, you forget to carry and say "115" instead of "125," then you not only did not do what you were disposed to do, but also did not do what you were disposed to regard as appropriate. This, I suggest, is sufficient to give content to the idea that you made a mistake. Conversely, if nothing interferes with your disposition and you respond with "125," you have done what you are disposed to regard as appropriate, and this makes it possible to think of your response as correct. In other words, modifying the dispositional view as I have proposed allows us to think of a disposition as the kind of thing which can be normatively conformed to or contravened in the way characteristic of rules and meaning. We can say that "125" normatively conforms to what you mean by "plus," or to the rule you have adopted for the use of the plus sign, and that "115" or "5" contravene what you mean, or the rule you have adopted, because these responses are not what you are disposed to regard as appropriate to the context. Because your disposition is infinite, in the sense that you are disposed to an infinite number of responses to all the possible "plus" questions, this approach does justice to the infinitary character of the state of your meaning addition by "plus." But it also does justice to the demand, implicit in the idea of a normative relation between meaning and use, that you represent all those uses as correct. That demand can seem impossible to satisfy, given that you cannot actually represent all the correct answers to possible queries. But you can be disposed, with respect to each of those answers, to regard it as correct, and my proposal works in part by exploiting this possibility.

If the proposal is accepted, then it yields an answer to the first and more important of the skeptic's two challenges. When the skeptic asks what it was about you in virtue of which you previously meant addition rather than quaddition, you can say that it was your being disposed to give the sum rather than the quum and, in so doing, to take yourself to be doing as you ought in the primitive sense. This answer satisfies

Kripke's normativity constraint on acceptable answers to the challenge because it allows us to make sense of the idea that you ought now to say "125" if you are to conform to your past meaning. For if you now say "5" you are failing to give what you were then disposed to regard as the primitively appropriate response, so that your response is incorrect relative to your past meaning. But it is important to keep in mind here that that the normative relation between your past meaning and your present use is different from the normative relation between past and present use ascribed in a claim to primitive normativity. That is, we need to distinguish between your taking "125" to be appropriate to the past sequence of your responses to "plus" questions, or "42" to fit the child's recitation of even numbers up to 40, and your taking "125" to be correct relative to the meaning of addition, or "42" to conform normatively to the add-two rule. First, while we saw that your meaning addition by "plus" in the past has no normative implications for your subsequent uses of "plus" and, more generally, that the normative significance of your meaning addition at any time is limited to its making intelligible ascriptions of correctness or incorrectness to your uses,· this is not true of the sequence of your past responses to "plus" questions. That you responded as you did to past questions makes it the case that it is appropriate, in the primitive sense, for you now to "go on" with "125" in answer to "68 + 57," just as the child's having recited the series of numbers makes it appropriate for her now to go on with "42."

Second, the appropriateness that you ascribe to "125" cannot be identified with correctness, at least in the sense that contrasts with incorrectness. When you take "125" to be appropriate with respect to your past uses, your attitude to other possible responses is not, at least in the first instance, that they are wrong, mistaken, or inappropriate, but simply that they are not appropriate, that is, that they lack the privileged normative status which you ascribe to your own response. Similarly, a child who takes "42" to be appropriate need not take another child who goes on with "43" to be making a mistake. Her attitude to her own response commits her to denying the appropriateness of "43," but this is not to say that she has to regard the second child as going on incorrectly rather than, say, uttering a word at random. Her conceiving the second child as mistaken requires something further, for example, the assumption that the second child is following the add-two rule or that she is disposed, like the first child, to regard "42" as appropriate. This point is essential to my account, which otherwise would be circular. Since I am using the notion of primitive normativity to make sense of the notion of meaning and claiming, in turn, that we need to ascribe meaning in order to make sense of the uses of expressions counting as correct as opposed to incorrect, it is crucial that we be able to understand the notion of primitive normativity in a way which does not depend on the notions of correctness or error.

We can see the contrast between primitive normativity and normative conformity to meaning by considering a case in which the two come apart. Suppose that the skeptical hypothesis is true of you now, so that, without realizing it, you now mean quaddition rather than addition. You are thus disposed, if queried about "68 + 57," to respond with "5" and to take yourself, in so doing, to be responding as you ought given your previous uses of the "+" sign. It follows that if nothing interferes with your disposition and you do, in fact, respond to the query with "5," then you are normatively according with what you mean by "+." That you mean quaddition makes it the case that "5" can be counted as a correct response, a point which we might also express by saying that "68 + 57 = 5" is true in your idiolect. If you had said "125," your answer would have counted as a mistake. But it is still the case that you are not responding as you ought in the primitive sense. Assuming that the "+" queries you have encountered so far have all involved numbers that are less than 57, the appropriate thing to say in the context created by your previous uses of "+" is not "5" but "125." Now it might be objected that if "5" is correct, or expresses a truth in your idiolect, then it must be the appropriate thing for you to say, or what you ought to say, regardless of your previous uses of "+." But it is appropriate only on the assumption that you ought to say what is true in your idiolect. And even if that assumption is accepted, the relevant normativity is different from, and less fundamental than, the primitive sense of normativity in which "125" is appropriate given your previous usage.[23]

I have proposed in this section that someone's meaning addition by "+" consists in her having a reliable disposition to use "+" in a certain way (specifically, in such a way as to respond to "+" questions with the sum) and, in so doing, to take that use to be primitively appropriate to the context.[24] This proposal extends to facts about grasping concepts and following rules more generally. Although in many contexts we do not count someone as having grasped a concept or rule unless he or she can articulate it explicitly, there is a less demanding notion which can be made out along the same lines I have suggested for meaning facts: what it is to grasp the add-two rule or the concept *green* is just to be reliably disposed to count by twos or to discriminate green things, with the proviso that one's counting or sorting behavior involves the consciousness of its primitive appropriateness to the context. It follows from this account that to the extent that the child in our initial example is competent in, say, counting by twos, she *eo ipso* grasps the add-two rule. And it is of a piece with that consequence that her utterance of "42" can be described as a case of her applying or following the add-two rule.[25] This might seem to call into question my characterization of the normativity revealed in these examples as "primitive." If, as in the case of the competent child, every case in which she takes herself to be going on as she ought is also a case in which she is applying the add-two rule, then, it might be argued, we cannot say that her taking herself to be going on as she ought does not depend on her recognizing that she is conforming to the add-two rule. On the contrary, it does depend on it, in the sense that if she does take herself to be going on as she ought, then it follows that she must be applying, and presumably also recognizing the applicability of, the add-two rule.

---

[23] The point here does not turn on "68 + 57 = 5" being true only in your idiolect rather than in the public language, but rather on the more general point that truth does not have intrinsically normative implications, that is, that the truth of an utterance does not makes it the case that you ought to assert it (or ought to refrain from asserting its negation) except in conjunction with explicitly normative principles, for example, moral or prudential principles prohibiting lies. This point is made by Boghossian, Wikforss, Hattiangadi, and others (see references in footnote 18).

[24] This account is not fully generalizable. If I stipulate that "+" is to mean quaddition and use it accordingly, then I might indeed be disposed to respond to "+" queries with

Here it is important to keep in mind that the account I am offering of meaning and rule-following is partially reductive in character. I am claiming that someone counts as grasping the add-two rule *in virtue of* being disposed to count by twos with the consciousness of primitive normativity, where both the disposition and the consciousness of the appropriateness of its actualizations are more basic than, and thus

---

the quum, but (unless my use of the "+" sign happens to be restricted to numbers less than 57) I will not take any of my responses to be primitively appropriate to the sign in light of its previous uses. In this case, the fact in which my meaning quaddition consists is a different kind of fact than in the addition case: it consists in my having made the stipulation and (perhaps) being disposed to use the term accordingly. But this case is atypical because, in contrast to most cases where we mean something by our words, someone cannot come to mean quaddition by an expression in any way other than explicit stipulation. The account which I have given is intended to apply only to cases—like those of "plus" and "green" in English—where we can come to use an expression meaningfully through the natural processes by which children typically acquire their first language.

[25] I disagree here with Boghossian's claim that what we ordinarily call rule-following requires an "intentional attempt to bring one's behavior in line with the dictates of [the] rule." See "The Rule-Following Considerations," p. 517. The conception I have offered represents a middle ground between that idea and the idea of "merely conforming" to a rule (*ibid.*, pp. 516–17).

explanatorily prior to, the fact of her grasping the rule. Now, it is the case for any reductive account that, if we accept it, we will have to accept a sense in which the reducing phenomena entail the facts to be reduced. If my account is correct, it will then follow that for someone who is reliable in counting by twos, the fact of her taking a given performance to be appropriate entails that she is applying the add-two rule and, presumably, taking her performance to conform to that rule. But if the idea of reductive explanation is intelligible at all, then there must remain a more fundamental sense in which the appropriateness she ascribes can be understood coherently as prior to and independent of her recognition that her performance conforms to the rule. So it can be accepted as a consequence of my account that claims to primitive normativity are never made without the recognition of corresponding rules, without undermining the primitive character of the normativity.[26]

IV

I turn now to a brief consideration of some objections that might be raised against my account. The first concerns my assertion, at the end of section II, that we not only make claims to primitive normativity, but are also entitled to do so. If we cannot appeal to rules or concepts in order to make such claims, it might be objected, what could possibly legitimize them? As I indicated at the end of section II, my commitment to the idea that such claims can be legitimate is based on our pretheoretical intuitions about particular cases. There is nothing to legitimize the child's claim to the appropriateness of "42" except our own intuition that "42" is appropriate. And that intuition, in turn, cannot be legitimized by appeal to any rule or principle, because it too reflects a claim to primitive normativity. But, on my proposal, this does not mean that such claims are illegitimate. On the contrary, if the account of rule-following sketched in the previous section is correct, it is only if we endorse the pretheoretical intuitions on which I am relying that we can make sense of there being justification in terms of rules in the first place. However, this appeal to intuition may seem unsatisfactory, in a way which can be brought out by imagining cases in which different people's intuitions conflict. Consider the "abnormal" child who goes on with "43." Her intuition that "43" is appropriate, let us suppose, is just as strong as the normal child's intuition about the appropriateness of "42." What entitles her to claim priority for her own intuition? Since the situations of the two children appear to be symmetrical, isn't it arbitrary for us to uphold the first child's claim while rejecting the second child's? Here my answer is simply to deny that the apparent symmetry is genuine. More specifically, while we might have the theoretically motivated thought that the two children's claims are equally legitimate, that thought is trumped by the pretheoretical intuition that "42" is appropriate and that "43" is not. That intuition is not undermined, as I see it, by the thought that in principle there could be intuitions which conflict with it, even if we know that the conflict would not be resolved by appeal to any neutral principle.[17]

---

[26] In fact, I am not inclined to accept this consequence, both because I think that children in the kinds of examples I have given can take certain of their performances to be appropriate without having acquired a reliable competence, hence without grasping a corresponding rule, and because I do not think that it holds in the case of aesthetic judgment. But it is not an essential commitment of my view that claims to primitive normativity can be made in isolation from the application of rules.

Here, and this is the second objection I want to consider, it might seem that I am simply missing Kripke's point, which is to undermine precisely the kind of pretheoretical intuition on which I am relying. Doesn't Kripke's argument boil down to the idea that since there is no justification for the appropriateness of "42" or "125," our uttering them must be a matter of going on blindly or making an "unjustified leap in the dark" (10)? So doesn't it exclude precisely what I am asserting, namely, that we can make claims to primitive normativity? I think, however, that this objection misrepresents Kripke's line of argument. It reads Kripke's skeptic as denying the legitimacy of claims to primitive normativity, whereas, as I see it, Kripke's skeptic and, presumably, Kripke himself do not consider the possibility of such claims. The skeptic does challenge our entitlement to claim that this or that response is appropriate in light of a preceding sequence of responses, but his challenge assumes just what the notion of primitive normativity is intended to deny, namely, that such a response can be appropriate only in virtue of its conformity to a rule exemplified by the preceding sequence. In section I, I drew attention to this assumption in connection with Kripke's quaddition example, but we also can see it in his discussion of an example closer to the one we have just been considering. "[Although] an intelligence tester may

---

[17] Bernard Williams makes a related point in the domain of ethical and political thought when he argues that our commitment to the basic principles of our own ethical outlook, for example, that every human being deserves equal consideration, is not undermined by the thought that there are competing outlooks which cannot be ruled out through arguments based on neutral premises. See Williams, "Philosophy as a Humanistic Discipline," *Philosophy*, LXXV, 4 (October 2000) 477–96. In effect, Williams takes our intuitive commitment to our own ethical outlook to trump whatever intuitions we might have favoring a symmetrical treatment of competing outlooks. I am proposing something analogous for the "outlook" exemplified by its being "42" rather than "43" which strikes us as the appropriate way of continuing the series.

suppose that there is only one possible continuation to the sequence 2, 4, 6, 8, ..., mathematical and philosophical sophisticates know that an indefinite number of rules...are compatible with any such finite initial segment. So if the tester urges me to respond, after 2, 4, 6, 8, ..., with *the* unique appropriate next number, the proper response is that no such unique number exists" (18). Here I want to draw attention to the "so" which begins the second sentence. Kripke is indeed denying the tester's entitlement to claim that "10" is the appropriate response, but his reason for denying it is that "2, 4, 6, 8..." might express the result of following any number of rules. He simply does not consider what I am proposing, namely, that we might legitimately regard "10" as appropriate to "2, 4, 6, 8..." in a sense which does not depend on its conforming to one rule rather than any other.[28] That proposal is of course open to challenge, but that would require a further stretch of argument which Kripke does not provide.

The third objection I want to consider concerns the adequacy of the positive account of meaning I offered in section III. My account relies on the idea that a subject can take her response, in a given context, to be appropriate, and that would seem to require that she entertain the thought *this is appropriate*. But Kripke's skeptical challenge is directed no less against the possibility of entertaining contentful thoughts than against the possibility of using expressions meaningfully. How then can I help myself to the idea of the subject's thinking of a response as appropriate? Isn't the subject's grasp of the concept *appropriate* vulnerable to the same kind of skeptical argument through which Kripke undermines the possibility of grasping *addition*? My answer is that while my account does presuppose that the subject has the capacity to entertain a concept of normative fit, which we might label as the concept *ought* or *appropriate*, there are grounds for treating it as a special case. In particular, I do not think that it is susceptible to the same kind of skeptical challenge which can be developed for *addition* or *green*, since our grasp of it does not seem to depend on a capacity to "go on" from the use of the corresponding word in particular cases. Now, it remains true that in taking for granted the consciousness of normativity, my approach falls short of a fully general account of intentional content. But this is just to say that the response it offers to Kripke's skeptical argument is only partially reductive. As I noted in the introduction, I aim to offer a middle way between a fully reductive naturalistic dispositionalism and the antireductionist view that meaning or content states are *sui generis*. My view has the advantage over traditional antireductionism that it is able to accommodate the quasi-dispositional character of meaning and rules, and so to tie meaning to our actual patterns of response. But it is nonreductionist to the extent that it requires us to accept at least one *sui generis* intentional attitude, namely, the attitude through which we take each of our actual responses to be appropriate to the circumstances.

---

[28] It might appear that I am simply being naive here, given that the segment "2, 4, 6, 8" figures in many sequences that are specifiable even without recourse to quus-like devices (637 such sequences are listed in N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, published electronically at www.research.att.com/~njas/sequences [2010]). Imagine a child, who, when asked by the tester to continue the sequence, says "30, 32, 126, 128, 8190, 8192, 13170," giving no explanation, but simply insisting—perhaps with surprise that the tester does not see it the same way—that this is the right way to go on. The tester might at first think that the child is saying numbers at random, only to find out later that the numbers coincide with the sequence of the nearest neighbors of the Mersenne primes. At this point, the tester will presumably conclude that the child is some kind of mathematical genius. It would seem absurd to claim that she was not "going on as she ought" from the initial segment. I still want to maintain, however, that she is not going on as she ought in the primitive sense with which I am concerned. Specifically, she is not entitled to claim that her utterance of "30" is appropriate in light of the sequence "2, 4, 6, 8," independently of any specification of a rule which that sequence instantiates. In effect, unlike the child who says "10," she owes the tester an explanation of why "30, 32" fits that initial sequence. And if, hard though this is to imagine, she lacks any disposition to continue the sequence with "10," then we can legitimately take there to be some defect in her mathematical abilities, in spite of her extraordinary gifts.

I want to conclude by indicating briefly the Kantian antecedents of my account. The notion of primitive normativity is drawn from the *Critique of Judgment* and, in particular, from Kant's account of judgments of beauty. According to Kant, a judgment of beauty makes a claim to universal agreement which does not, as he puts it, "rest on any concept" (§8, 5:215).[29] When you respond to an object in such a way as to experience it as beautiful, you respond to it in a way which involves your taking it that everyone ought to respond to the object in the same way you do. But the claim to agreement implicit in your response is not based on your recognition of the object as falling under a concept or satisfying a criterion. In my terms, then, you make a claim to primitive normativity. Your response is analogous to the child's saying "42," or your saying "10" in the example of the intelligence test: it involves your taking yourself to respond as you ought, but where the "ought" does not depend on the assumption that you are bound by an antecedently specified rule.

Aesthetic judgments are, of course, a special case, so one might wonder about the propriety of drawing on Kant's model of aesthetic judgment to make a point about rules and meaning in the nonaesthetic domain. But Kant makes clear that he takes his account of aesthetic judgment to have broader implications. The principle of aesthetic judgment, he says, is the "principle of the faculty of judgment in general" (5:286), and the faculty of judgment has a significance for Kant that goes far beyond the aesthetic domain. Judgment, in its most general definition, is the faculty for "thinking the particular as contained under the universal" (Introduction IV, 5:179); that is, it is our capacity for subsuming particular things under rules or concepts. Now, judgment in the first Critique is treated either as subordinate to the faculty of understanding or as identical with it. The rules or concepts under which judgment subsumes particulars are thought of as already grasped by the understanding, and the role of judgment is simply to apply these concepts to the particular instances that satisfy them.

---

[29] All references to Kant are either to the *Critique of Judgment* or to the associated "First Introduction." I cite the section number followed by volume and page number in Immanuel Kant, *Gesammelte Schriften* (Berlin: De Gruyter, 1902–); these page numbers are typically found in the margins of recent English translations of Kant. Translations are my own.

But in the third Critique Kant introduces the idea that judgment can function independently of the understanding in situations where the relevant rules or concepts are not already specified. To exercise judgment in this independent way is to judge particulars to be contained under rules or concepts which are, so to speak, not already in the understanding but rather made possible by those acts of judging, themselves. Aesthetic judgment is the paradigmatic exercise of judgment in this autonomous capacity. It is in aesthetic judgment, Kant says, that the faculty of judgment "reveals itself as a faculty which has its own special principle" and thus as claiming a place of its own, beside understanding and reason, in the "general critique of the higher cognitive faculties" (First Introduction XI, 20:244).

Kant's claim that aesthetic judging reveals the faculty of judgment as an independent capacity obviously deserves more discussion than I can provide here. But I believe that what lies behind it is the idea that the apparently anomalous character of aesthetic judgments—that they involve normative claims to agreement which do not rest on rules or concepts—indicates something which is required for the possibility of rules or concepts in the first place. For I understand Kant to hold that the possibility of grasping rules or concepts, which in turn can be applied in particular cases, depends on the possibility of our taking our responses to those particular cases to be normative in a way which does not rest on the assumption of rules or concepts.[30] That is the point which I have been trying to make by applying the notion of primitive normativity against Kripke's skepticism. Kripke's skepticism, I have argued, is motivated by an unexamined assumption about the

---

[30] For more discussion of this point, see my "Lawfulness without a Law: Kant on the Free Play of Imagination and Understanding," *Philosophical Topics*, xxv, 1 (1997). 37–81; and "Aesthetic Judgment and Perceptual Normativity," *Inquiry*, XLIX, 5 (October 2006) 403–37.

COMMENTS AND CRITICISM

IS DIRECT REFERENCE THEORY INCOMPATIBLE WITH PHYSICALISM?[*]

The answer, one would naturally think, is "No, of course not. Direct reference is a semantic theory, and physicalism is a metaphysical doctrine. To think that they are incompatible would involve some kind of obvious confusion." But in a recent article, Thomas Hofweber presents an argument suggesting that, despite appearances to the contrary, physicalism is incompatible with the theory of direct reference.[1] Hofweber wants to be neutral regarding which doctrine one ought to give up.

We think Hofweber's argument is unsuccessful; direct reference and physicalism are compatible. In this paper we will explain why. We will show that the apparent conflict between direct reference and physicalism is due to a latent inconsistency in the way Hofweber defines a couple of technical notions, each of which plays an important role in the argument we want to criticize. Specifically, the way Hofweber defines what he calls a "local" property is such that one will think there are local properties only if one is a haecceitist. Furthermore, the way Hofweber defines the notion of a physical property is such that one will be a physicalist only if one is an anti-haecceitist. The inconsistency, then, is due to the fact that different parts of Hofweber's argument presuppose contradictory views about how *de re* facts—facts about particular objects—relate to qualitative facts.

In section I we present Hofweber's argument. In section II we recommend three different ways of rejecting it and reply to some possible objections. We then observe that if Hofweber's argument is sound, some of his independent commitments logically compel him to reject physicalism. Contrary to what he says, Hofweber cannot be neutral about which of direct reference and physicalism is false. Throughout, we try to disentangle as best we can semantical claims from metaphysical claims. One can view the discussion that follows as a case study in the interaction between semantics and metaphysics.

---

[*] We thank Alex Byrne, Sally Haslanger, Dilip Ninan, Bob Stalnaker, and Steve Yablo for comments and questions. Special thanks to Thomas Hofweber for email correspondence.

[1] Thomas Hofweber, "Supervenience and Object-Dependent Properties," this JOURNAL, CII, 1 (January 2005): 5–32.

0022-362X/11/0805/255–268

I. A BRIEF SUMMARY OF HOFWEBER'S ARGUMENT

We begin by presenting the argument we want to criticize.

(P1) If direct reference theory is true, then object-dependent properties exist.
(P2) If physicalism is true, then object-dependent properties do not exist.
(C) Therefore, direct reference and physicalism cannot both be true.

Hofweber does not say which of direct reference and physicalism he thinks is false. The argument is thought to generate a puzzle.[2]

One preliminary remark before we begin reconstructing the case for (P1): Hofweber formulates his argument in terms of properties because he takes the canonical statement of physicalism to be in terms of properties. But he assures us that nothing rests on the presupposition that there are properties. One can reformulate Hofweber's argument in terms of facts, propositions, or any other kind of entity that one thinks plays the role that physical properties traditionally play—namely, an ontologically fundamental supervenience-base on which nonfundamental entities of the same kind supervene. We accept this way of setting things up without protest for ease of exposition; we will present Hofweber's argument, and our reply, in a way that presupposes the existence of properties. Among the properties we, and Hofweber, take there to be are *thisnesses*, for example, the property of being Bill Clinton. Each thisness can be exemplified by no more than one object, so we distinguish them from general properties, which can be shared.

We now turn to the case for (P1). If direct reference theory is true, then proper names, pronouns, and demonstratives semantically contribute their referents directly, not by way of expressing a sense or individual concept or whatever. Take any simple sentence, *s*, in which a name, *n*, occurs in subject position. The direct reference theorist will say that its truth-value, in the actual world and in counterfactual situations, will be determined as a function of *n*'s referent.[3] So, if one substitutes for *n* a coreferring name, *n*′, the truth-value of the new sentence, *s*′, one thus obtains will not differ from the truth-value of *s*, either in the actual world or in counterfactual situations. Since *s* and *s*′ differ only with respect to what occurs in their subject positions and the semantic contributions of their subject terms do not differ, *s* and *s*′ express the same proposition. Furthermore, says Hofweber, that proposition will not exist in worlds where *n*'s referent (also the referent of *n*′) does not exist. So, Hofweber concludes, if direct reference theory is true, then the semantic values of simple subject-predicate sentences will depend for their identity and for their existence on the objects to which their subject terms refer.

---

[2] In correspondence, Hofweber has said that the way we characterize the first premise of his argument is not quite right. The claim is not, according to him, that direct reference theory is true only if there are object-dependent properties. The claim is, rather, that if direct reference theory is true, then physical objects have object-dependent properties. The key issue, for Hofweber, is not what kinds of property a direct reference theorist must think there are; the key issue is what kinds of property a direct reference theorist must think physical objects have. But if a direct reference theorist must think that physical objects have object-dependent properties, then she must also think that there are object-dependent properties to be had. So the way we characterize Hofweber's first premise is, strictly, weaker than the way he would like to characterize it. We believe, therefore, that our way of presenting the dialectic does not unfairly distort the issue at stake here. If one thinks that something of interest turns on this point, then one should feel free to reinterpret our discussion in terms of having properties; reinterpreting what follows in this way will not affect our primary response to Hofweber's argument, which we explain in sections II.2 and II.3. We do not pursue this line of thought further because it raises questions which are not directly relevant to the problems we want to address here. We thank Hofweber for discussion.

Now, assume for the sake of argument that the semantic value of a predicate is a property. What the direct reference theorist says about sentence meanings, she will say about predicate meanings. The semantic value of a predicate, *P*, in which a name, *n*, occurs as a part is similarly determined as a function of *n*'s referent. So, Hofweber believes, the direct reference theorist will say that if one substitutes for *n* in *P* a coreferring name, *n*′, then the new predicate one obtains will express the very same property. Furthermore, Hofweber claims that that property will not exist in worlds where *n*'s referent (also the referent of *n*′) does not exist. If, however, one substitutes for *n* in *P* a noncoreferring name, then the predicate one obtains will express a different property. Therefore, the truth of, say, 'Hillary Clinton is Bill Clinton's wife' demands that there be a property—the property of being Bill Clinton's wife—whose identity and existence depends on a particular object, namely, Bill. Since any reasonable direct reference theorist will admit that Hillary is Bill's wife, she will have to acknowledge the existence of object-dependent properties. So, if one affirms the theory of direct reference and accepts as true that, for example, Hillary is Bill's wife, then one will have to say that there are object-dependent properties. Since no reasonable direct reference theorist would want to deny that Hillary is Bill's wife, we can blamelessly abbreviate the conclusion of this argument and simply say that direct reference is true only if there are object-dependent properties.

---

[3] Someone who thinks that names have rigid descriptive content will, of course, agree with this idea, but what separates direct reference from rigid descriptivism is the further thought that names are not synonymous with any descriptions.

The argument for (P2)—physicalism is true only if there are no object-dependent properties—proceeds as follows. First, Hofweber observes that physicalism does not deny that there are nonphysical properties; it says that the nonphysical properties are determined by the physical properties. One often wonders what the physical/nonphysical distinction amounts to, and what it is for one family of properties to determine another family. We will discuss each question in turn.

Hofweber takes a physical property to be a property that plays a role in affecting the behavior of fundamental physical objects. Examples of such properties are having a specific mass, having a particular spin, and being positively charged. It seems highly unlikely, Hofweber says, that object-dependent properties play a role in affecting the behavior of fundamental physical objects, over and above the role that qualitative properties (mass, spin, charge, and so on) play. So, Hofweber concludes that all and only the kinds of properties that more than one thing can have, and that a complete physics will catalog, are physical.

> To see which properties are the physical ones we have to look back at what motivated the idea of physicalism, the idea that all the physical determines the rest. The idea, again simply put, was that all physical objects are made up from smallest stuff, and this smallest stuff has basic properties and relations which together are sufficient to determine its behavior....The physical properties are thus the smallest class of properties that are sufficient to determine the behavior of the smallest stuff....And it is here that the strangeness and irrelevance of object-dependent properties becomes apparent at the physical level. Even if the smallest stuff has object-dependent properties, these properties are completely irrelevant for their behavior.
>
> Take a simple example, like a particular electron, call it Elly. Elly has a number of general properties which are not object dependent. It has a charge, mass, speed, and so on....The property of having a certain charge will affect its behavior in one way, and the property of having a certain mass will affect it in another. But the property of being Elly has nothing to do with its behavior. It will behave exactly like any other electron that has the same mass, charge, and so on.[4]

Two points deserve emphasis. First, if Hofweber is right, it then follows that only qualitative properties are physical; the property of being Elly—though had by a physical object—is nonphysical. Keep this point in mind; it will play an important role in the rest of Hofweber's argument. In the next section, we will challenge the idea that the property of being Elly is nonphysical. Second, the idea that motivates physicalism is that the facts concerning the smallest stuff—what properties it has and how it is distributed in space-time—suffice to determine how things behave. Following Hofweber, we take this idea to be constitutive of physicalism.

---

[4] *Ibid.*, p. 18.

What is it for one thing to determine another? To analyze the concept of determination, theorists often appeal to the notion of supervenience. *x* determines *y* (if and?) only if *y* supervenes on *x*. Appeal to supervenience was thought to offer a physicalistically kosher alternative to reduction. Whether this is right is controversial. Fortunately, we can ignore the controversy. All that matters for our purpose is the particular supervenience thesis that Hofweber recommends. It involves a kind of property that Hofweber calls 'local'. He defines a local property as follows: "A property, physical or nonphysical, is local iff a physical object inside the solar system has that property and that object would still have this property even if the universe outside the solar system were different, but the same inside the solar system (over time)."[5] Locality is, therefore, a kind of modal flexibility;[6] whether an object has a local property in a world depends only on what is happening inside our solar system at that world. With this in mind, we are now able to specify the relevant supervenience principle that, according to Hofweber, characterizes physicalism. The local nonphysical properties supervene on the local physical properties iff

> (LLS) Necessarily, for every local nonphysical property *N* that a physical object *o* has, there is a local physical property *P* that *o* has such that: necessarily, if something has *P*, then it has *N*.

In order to bypass well-known objections which suggest that so-called 'strong' supervenience is too weak for the physicalist's purpose, Hofweber invites us to accept (LLS) as a necessary condition for one thing to determine another. For the sake of argument, we accept Hofweber's invitation.

If physicalism is true, then (LLS) is true. But, if (LLS) is true, can there be object-dependent properties? Hofweber argues that the answer is no. His argument begins with the claim that object-dependent properties, if they exist at all, are local properties. An object would not cease to have the property of being Clinton, for example, if things outside the solar system were different but things inside were the same. Hofweber takes this claim to follow trivially from the claim that being Clinton is an object-dependent property.

---

[5] *Ibid.*, p. 14.

[6] One should not confuse Hofweber's notion of a local property with other technical notions that go by similar names; Hofweber does not intend to analyze any notions in physics or its philosophy

> Since the property of being Clinton is merely the property of being identical with a certain object, changing around which objects there are outside of the solar system and which properties they have has no effect on which objects there are inside the solar system. Clinton would still be there and he would still have the property of being identical to Clinton. This might seem trivial, and it is. But it is only trivial if we assume that the property of being Clinton is an object-dependent property. If it were the property of being the *F*, for some purely general *F*, as certain descriptive theories would hold, then this is not trivial any more, and is in fact false. If the property of being Clinton is the property of being the *F* then we could change the world outside around in such a way that outside of the solar system we added another *F*. Thus Clinton would not be the (one and only) *F* any more. But under the assumption that the property of being Clinton is an object-dependent property it indeed follows trivially that it is a local property.[7]

Now, suppose for a moment that (LLS) is true. Since object-dependent properties are (we will grant for the moment) both local and nonphysical, they must supervene on local physical properties. Additionally, some object-dependent properties are *uniqueness properties*—at most, a single object in any world exhibits the property of being Clinton—so they must supervene on properties that are uniquely co-instantiated. (If being Clinton were determined by a supervenience-base that was not itself uniquely realized, then, by definition, there would be a possible world in which the supervenience-base was realized in multiple regions of that world. So, in that world, the property of being Clinton would be had by more than one object. But there is no such world, because being Clinton is a uniqueness property.) "However," says Hofweber, "no physical property is both a local property and a uniqueness property."[8] So, if (LLS) is true, there are no object-dependent properties.

Let us ask why Hofweber thinks physical properties cannot be local uniqueness properties. Think about the possible ways in which *P*, say, might be a uniqueness property. Either *P* is object dependent, in which case it is not physical, or *P* is what Hofweber calls a logical uniqueness property, like being the tall, dark, and handsome former senator from Illinois and current commander-in-chief, which presupposes uniqueness in virtue of the way it is described. But, Hofweber observes, logical uniqueness properties cannot be local. To see why, simply assume that *P* is the property of being the *F*. Now, suppose for the sake of argument that being the *F* is a local property. It would then follow that an object, *o*, which, let us say, is the *F*, would still be the *F* even if the universe beyond our solar system were different. But God could simply create another object in the Andromeda Galaxy that exhibits *F*-ness, in which case *o* would no longer be the (unique) *F*. So, being the *F* is not a local property; altering the world outside our solar system in a particular way would cause an object that has the property to cease having it. It seems, then, that none of the possible ways in which *P* can be a uniqueness property allow for it to be both local and physical. So it seems that if (LLS) is true, then there are no object-dependent properties.

---

[7] *Ibid.*, p 16

[8] *Ibid.*, p. 17.

II. CONSILIENCE

We think both premises of Hofweber's argument can be rejected on reasonable grounds. Our discussion begins with the first premise, namely, that direct reference is true only if there are object-dependent properties.

II.1. Direct Reference and Object Dependence. We should all agree that if Bill Clinton never existed, his wife, Hillary, still could have. Of course, the direct reference theorist will say that we would not be able to refer to her as 'Bill Clinton's wife'—because we would not be able to refer to Bill Clinton—but even the direct reference theorist can agree that Hillary's existence is not inseparably bound up with Bill's. The sort of direct reference theorist we have in mind might think something very similar about the property of being Bill Clinton. Suppose the direct reference theorist takes a property to be an abstract way an individual might be, and thinks that all the ways an individual might be exist, but only some of them are realized.[9] (Direct reference theorists are perfectly happy to grant that there are abstract ways a world might be and that all of them exist, but only one is instantiated.[10] So we see no reason for these theorists to balk at the idea that there are uninstantiated ways that lesser individuals might be.) Now, suppose that Bill Clinton never existed. On the view we have in mind here, the property of being Bill Clinton would still exist; it just would not have an instance, and we would not be able to pick it out as a function of Bill. If the direct reference theorist thinks of properties in this way, then she will think that whether the property of being Bill Clinton has an instance depends on whether Bill exists, but the identity and existence of the property of being Bill Clinton itself does not depend on Bill's existing.

---

[9] Alvin Plantinga's theory of essences is very like the theory of properties we have in mind above. Thanks to Bob Stalnaker for discussion here.

[10] See, for example, Scott Soames, "Actuality: Actually," *Aristotelian Society Supplementary Volume*, LXXXI, 1 (2007): 251–77. See also Robert Stalnaker, "Possible Worlds," *Noûs*, X, 1 (March 1976): 65–75.

One would think that 'is Bill Clinton' semantically contributes an object-dependent property only if one thought that 'Bill Clinton' semantically contributes an object which is literally a component of the property that 'is Bill Clinton' contributes. But that thought involves more than direct reference theory; it involves a particular view of properties as structured entities, and a direct reference theorist need not think of properties in that way, just as she need not think of propositions as having structure.

The theory of direct reference is sometimes characterized as the view that sentences of the form '*a* is *F*' semantically express singular propositions. Some theorists go on to define the notion of a singular proposition in a way that presupposes a commitment to structured contents.

> A singular proposition is a proposition in which at least one individual or object that the proposition is about occurs directly as a constituent, and the proposition is about that individual by virtue of directly including it, rather than a concept by which the individual is represented (determined, denoted).[11]

David Kaplan contrasts direct reference theory with Frege's theory of sense and reference in terms of a referent occurring directly as a "propositional component," rather than as something determined by a propositional component.[12] These ways of presenting direct reference theory encourage the mistake of thinking that direct reference theory entails that the semantic values of complex expressions involving proper names—such as the predicate 'is Bill Clinton's wife'—are complex entities which have as a proper part the referent of the name. It is this mistake that underpins Hofweber's first premise.

Direct reference theory is simply the view that the semantic value of a complex expression in which a name, *n*, occurs is determined as a function of *n*'s referent. One can coherently ally this view with a view of sentence and predicate meanings that assigns unstructured entities as their semantic values. This would allow one to say that the semantic value of 'is Bill Clinton's wife' or 'is Bill Clinton' would still exist even in worlds where Bill Clinton does not; we just would not be able to pick it out as a function of Bill himself. We conclude, therefore, that direct reference does not imply object dependence.

---

[11] Nathan Salmon, "Trans-World Identification and Stipulation," *Philosophical Studies*, LXXXIV, 2-3 (December 1996): 203–23, at p. 208.

[12] David Kaplan, "Demonstratives," in Joseph Almog, John Perry, and Howard Wettstein, eds., *Themes from Kaplan* (New York: Oxford, 1989), pp. 481–563.

It is open to Hofweber to retreat to the weaker but still interesting claim that direct reference theory plus structured contents is incompatible with physicalism. It would, after all, be very odd if one could not consistently be a physicalist, a direct reference theorist, and take predicate meanings to be structured. *Prima facie*, physicalism has nothing to do with what sorts of representational tools one uses to compositionally specify the contents of natural language expressions. Additionally, many direct reference theorists do, in fact, think contents are structured. We believe that even if Hofweber opts for this response, the argument is unsound. Direct reference theorists who go in for structured predicate meanings can deny (P2). In sections II.2 and II.3 we will explain how. Our explanation will presuppose that direct reference does, contrary to fact, imply object dependence. So it will be available to any direct reference theorist, regardless of her views about the kind of thing that should occupy the role of predicate meanings.

Remember that the case for (P2) rests on at least two claims.

(P2a) Object-dependent properties are local uniqueness properties.
(P2b) Object-dependent properties are not physical.

Physicalism and (P2a) together imply that if object-dependent properties are not physical, then they must supervene on physical local uniqueness properties. In light of this and (P2b), it follows that object-dependent properties must supervene on physical properties that are both local and unique. But since physical properties are qualitative (or general), on Hofweber's view, they are not local uniqueness properties.

We will argue that one will accept (P2a) only if one is a haecceitist. Similarly, we think one will accept (P2b) only if one is an anti-haecceitist. Nobody who has her views straight about haecceitism should accept both premises, so nobody who has her views straight about haecceitism should accept Hofweber's dramatic conclusion. What Hofweber's argument really shows, we think, is not that direct reference is incompatible with physicalism, but that haecceitism is incompatible with anti-haecceitism. That is a conclusion we are ready to accept.

II.2. Object Dependence and Locality. Consider the property of being Bill Clinton, and assume that it is object dependent. Now, according to Hofweber, it will seem trivial that being Clinton is a local uniqueness property. "But it is only trivial if we assume that the property of being Clinton is an object-dependent property. If it were the property of being the *F*, for some purely general *F*...then this is not trivial any more, and is in fact false."[13] The reason it would be false, according to Hofweber, is that God would then be able to change things beyond our solar system, while keeping things exactly as they are within it, in such a way that the property of being Clinton would no longer be realized. But in order for object dependence to guarantee locality and uniqueness, one has to assume something stronger than the thought that the property of being *o* is distinct from the property of being the *F*; one has to assume that whether the *F* exists does not determine whether *o* exists. Otherwise, God would be able to annihilate *o* and the corresponding thisness simply by willing that there be an *F* in the Andromeda Galaxy. But this amounts to the assumption that whether an object exists at a world, and thus whether that world is one in which the corresponding object-dependent properties exist, is not determined by whether something exhibits some set of purely qualitative properties. Now, there is a name for such a view about the relationship between particular objects and qualitative properties; it is 'haecceitism'. The haecceitist thinks that *suchness*, or qualitative character, does not determine thisness; facts about which particular things there happen to be do not counterfactually covary with facts about which qualitative properties happen to be realized. There are possible worlds that are qualitative duplicates—they are exactly alike with respect to the distribution of qualitative properties—but which differ with respect to facts involving particular objects. The anti-haecceitist, in contrast, thinks suchness determines thisness; God has only to fix the qualitative facts and he thereby fixes the *de re* facts.

---

[13] Hofweber, *op. cit.*, p. 16.

A direct reference theorist can coherently accept anti-haecceitism. Direct reference is a thesis about the semantic relationship between proper names, pronouns, and demonstratives on the one hand and definite descriptions on the other; it says that expressions of the first sort are not semantically equivalent to expressions of the second. Anti-haecceitism is a doctrine about the metaphysical relationship between particulars and qualitative properties; it says that facts involving things of the first sort depend on and are determined by facts involving things of the second. It is perfectly consistent with direct reference theory that whether a particular object exists at a world, and thus is available to be the semantic value of a directly referring term at that world, is determined entirely by the distribution of qualitative properties at that world. It would follow, then, that a directly referring term, *t*, has a semantic value at a world iff a certain description, *d*, aptly characterizes a part of that world. This does not contradict the theory of direct reference, however, because *d* need not be semantically equivalent to *t*. So one can coherently think that the property of being Clinton is object dependent (no being Clinton without Clinton) and that *o* is Clinton iff some purely qualitative properties are had by *o*, and thus that some purely general description aptly characterizes *o*. And, if one accepts these views, then the property of being Clinton will not be a local uniqueness property. We conclude, therefore, that (P2a) presupposes haecceitism.

To summarize: object dependence does not imply that there are local uniqueness properties. If one assumes a specific view about the nature of particularity and accepts the theory of direct reference, then one will have to admit that there are local uniqueness properties. But neither direct reference theory nor the thought that being *o* is object dependent commits one to any such view about the metaphysics of particulars.

II.3. 'Physical', Physicalism, and Object Dependence. We have said that (P2) rests in part on (P2b). We think a direct reference theorist can—consistently with physicalism—deny (P2b). What motivates (P2b) is the example involving Elly the electron, so we begin our response by returning to it.

Consider Elly. Hofweber says that the property of having a negative charge affects it so that it behaves one way, and the property of having a specific mass affects it so that it behaves another way. "But the property of being Elly has nothing to do with its behavior." Since the property of being Elly does not determine the behavior of any fundamental physical object, it must not, Hofweber concludes, be physical.

The direct reference theorist should, we think, deny that the property of being Elly does not play a role in affecting Elly's behavior. What affects *Elly* so that she behaves in this or that way is not merely that some particle or other has a negative charge, a specific mass, and so on. After all, any number of electrons can have these particular qualitative properties without *Elly* behaving in this or that way. What causes Elly to behave in the way she does is that Elly has a negative charge; Elly has a specific mass; and so on. Something's having a negative charge, a mass of x grams, a velocity of y meters per second, and so on is not sufficient for Elly to behave in this or that way unless that thing also has the property of being Elly. If those very same qualitative properties are had by something that has the property of being Electra, and not had by something that has the property of being Elly, then Electra and not Elly would behave in this and that way. So, *pace* Hofweber, the property of being Elly has a lot to do with Elly's behavior; only when a thing has that property will its qualitative properties suffice to determine Elly's behavior. Why would one think otherwise? One would think otherwise if one thought that the *de re* facts—the facts about how Elly, for instance, behaves—are determined by the qualitative facts.

There are two contrasting pictures of how the *de re* facts relate to the qualitative facts, and Hofweber's argument, we think, presupposes one of them. By asserting that the property of being Elly has nothing to do with Elly's behavior, Hofweber tacitly assumes that a qualitative characterization of a thing's behavior is a complete characterization of its behavior. But that is just to assume the anti-haecceitist's picture of how the *de re* facts relate to the qualitative facts, and it is this picture that a direct reference theorist can reasonably deny.

In responding to Hofweber's argument, the direct reference theorist can admit that there are object-dependent properties; she can even acknowledge that physical properties are causally relevant, not merely epiphenomenal; but she will have to say that object-dependent properties are causally relevant. Being Elly, for example, partly determines Elly's behavior. Something's having that property is necessary for Elly to behave in the way that she does. So, the sort of theorist we have in mind can go on to say that being object dependent is compatible with being physical and, therefore, that physicalism is compatible with direct reference. Of course, as the discussion above suggests, this line of thought rests on a prior commitment to haecceitism. We do not find this commitment unattractive, though, nor do many well-known advocates of direct reference theory.[14]

Hofweber's dramatic conclusion—that the theory of direct reference and physicalism are logically incompatible—tacitly involves a commitment to the thought that physicalism is true only if anti-haecceitism is. Recall, to begin with, that Hofweber characterizes the notion of physicality in terms of causal relevance; a property is a physical property, he says, only if it plays a role in affecting how fundamental physical objects behave. Hofweber then appeals to the example involving Elly to motivate the idea that only qualitative properties are causally relevant. If all of this is right, then only qualitative properties will count as physical, and physicalism will imply anti-haecceitism, because the physical properties will determine all of the properties only if the qualitative properties determine all of the properties. But physicalism does not imply anti-haecceitism, because a physicalist can consistently think that being Elly is both causally relevant to Elly's behavior and irreducible to any set of qualitative properties. We conclude, then, that Hofweber's argument fails to establish its conclusion.

Hofweber is aware of a connection between his argument for the incoherence of direct reference and physicalism on the one hand

---

[14] For expressions of sympathy, see Kaplan, "How to Russell a Frege-Church," this JOURNAL, LXXII, 19 (Nov. 6, 1975): 716–29; Stalnaker, "Indexical Belief," reprinted in *Context and Content* (New York: Oxford, 1999), pp. 130–49, and Salmon, *op. cit.*

and haecceitism on the other. Drawing on some of the concepts we introduced in section I—in particular, the concept of locality—Hofweber argues that haecceitism is true. For our purpose here we need not delve into the specifics of this argument. What matters for our purpose—and what we now want to focus on—is that in light of the way Hofweber defines the notion of a physical property, physicalism is true only if anti-haecceitism is. Given Hofweber's independent commitment to haecceitism and his conception of the physical, Hofweber must think that physicalism is false. He cannot be neutral about the dilemma he takes himself to have identified.

One might criticize our response to (P2b) as follows. Physicists do not characterize phenomena in object-dependent terms. In particular, physicists do not care about *Elly's* behaving in this or that way; they only care about the way particles having such-and-such mass, spin, charge, and so on interact with other particles having such-and-such mass, spin, charge, and so on. The only sort of behavior that is theoretically significant to physicists is behavior understood in qualitative terms. More generally, the aim of physics is to specify the laws governing how physical objects interact. Now, it is highly unlikely that such laws will involve reference to particular physical objects, so it is highly unlikely that these laws will invoke object-dependent properties. But a property is physical only if it figures in the sort of laws that a complete physics will specify, so it is highly unlikely that object-dependent properties are physical.[15]

In section I we said, following Hofweber, that a property is physical only if it is causally relevant to the behavior of fundamental physical objects. Our critic's objection, however, rests on the thought that a property is physical only if it figures in the laws governing how fundamental physical objects interact. If one has a kind of nomological view of causal relevance, according to which a property, *P*, is causally relevant to an event, *e*, iff *P* figures in a law that governs the kind of event of which *e* is an instance, then these two criteria of physicality—Hofweber's and our critic's—will amount to the same thing. But physicalism does not force one to adopt a nomological view about causal relevance, so the two ways of characterizing the notion of a physical property can come apart. We see no reason why a physicalist should prefer our critic's notion of a physical property to Hofweber's. Physicalism, as we said, is simply the idea that the character of the most fundamental physical objects suffices to determine everything—that macro-level goings on do not float free from micro-level physical goings on. Nobody thinks that the laws by themselves suffice to determine how particular physical objects behave; the laws plus particular facts about the relevant objects—that they have such-and-such a mass, spin, charge, and so on—determine how physical objects behave. For the haecceitist, the property of being Elly is, in this respect, just like the property of having such-and-such a mass, spin, charge, and so on. Without appealing to it, one will not be able to specify a set of properties that suffice to determine the behavior of particular objects. Thus the haecceitist will think that one can respect the simple idea behind physicalism only if one acknowledges that the property of being Elly—indeed, any identity property of a fundamental physical object—is itself physical.

---

[15] Hofweber suggested a response similar to this one (personal communication).

MAHRAD ALMOTAHARI
DAMIEN ROCHFORD

Massachusetts Institute of Technology

REVIEW ESSAYS

Origins of Objectivity. TYLER BURGE. New York: Oxford University Press. xx + 636 p. Cloth $40.00.

Burge's official target in this book is what he calls "Individual Representationalism": the view that "an individual cannot empirically and objectively represent an ordinary macro-physical subject matter unless the individual has resources that can represent some constitutive conditions for such representation" (13). This formulation allows for variation in what one takes the "constitutive conditions" of representation to be, and says that whatever they are, the subject has to represent at least one of them. The formulation of Individual Representationalism has to be quite general, though, since it has to cover Burge's four main opponents: Strawson, Evans, Davidson, and Quine, who on the face of it all hold rather different views.

Burge's opponents argue variously that spatiotemporal thought, or grasp of criteria of identity for objects, or conscious experience, or triangulation with other subjects, is required for objectivity in propositional thought. Burge's response is always that the objectivity of propositional thought does not depend on those conditions, but only on its connections to the representations postulated by information-processing perceptual psychology. The central issue raised by the book is whether the objectivity of linguistic or conceptual thought can be explained entirely in terms of the connection of thought to the perceptual representations postulated by perceptual psychology.

Burge accuses his opponents of "hyper-intellectualizing" objective content. His positive view is that the origin of objectivity is provided by perceptual representations. Perceptual representations have their contents determined by facts about the environment the organism occupies, its biological needs, and its capacity to perceive constancies in its surroundings. All that is required for objective representation is, fundamentally, one's perception of the things and properties that causally affect one, in the context of one's biological needs and a capacity to perceive perceptual constancies.

His opponents were concerned primarily with the analysis of linguistic or conceptual thought. Burge says that there is a distinction between that kind of content and the representational content of perception ("Think of a map," he says (540)). He says that explaining the distinction between conceptual and perceptual content deserves

0022-362X/11/0805/269–285

"a project in itself" (539). Burge says that linguistic or conceptual thought is grounded in perceptual representation. But he writes that he has tried to frame his discussion "so as not to rely in any fundamental way on my views regarding the relation between perception and propositional thought" (537). He does give a brief sketch of how he thinks the grounding of propositional thought in perception might go (537–44). But the main focus of his positive account of content is perceptual representation. His opponents were talking about propositional thought; he is talking about perception. His task is to give an account of the objectivity of perceptual representation that will show how it could be enough to ground the objectivity of propositional thought.

The main novelty in Burge's account of perception is his emphasis on the idea of perceptual constancy. In much of the philosophical and psychological literature on perceptual constancy, the interest is quite a basic one in saying what the phenomenon *is*. One way of thinking about, for example, color constancy, is that it is a matter of the subject being able to extract the same perceptual representation of the absolute value of the color of a surface through, say, variation in the illuminant. Another way of thinking of color constancy is that it is a matter of being able to sustain a perceptual representation of the sameness or difference in color of two surfaces through, say, variation in the illuminant. These are different phenomena. The first will be a coarser measure of constancy than the second in the cases in which the visual system is more sensitive to the relations between perceived properties than it is to the absolute values of those properties. The second, "relational color constancy," is consistent with the absolute color value assigned to both surfaces by vision actually changing with variation in the illuminant.[1] Burge does not address this kind of question. Some of his formulations suggest that he is thinking in terms of sameness of the absolute color value perceptually represented: "the colors attributed in exercises of perceptual constancies are the same, when they are the same" (412). Sometimes he seems to be thinking of perception as representing the sameness or difference between, for example, colors, in exercises of perceptual constancy, when he talks of "capacities systematically to represent a given particular or attribute as the same despite significant variations in proximal stimulation" (274; cf. 114). Empirically, it looks as though cognition plays a role in color constancy: memory of the colors of fruits, for example, seems to play a role in color constancy.[2] I do not think Burge means to be rejecting this kind of finding when he says, "The constancies emphatically do not depend on knowledge or conceptual understanding" (412). He does not address the question directly.

---

[1] For a representative discussion, and further references, see David H. Foster, "Does Color Constancy Exist?" *Trends in Cognitive Sciences*, VII, 10 (October 2003): 439–43. For an excellent overview, see Jonathan Cohen, "Perceptual Constancy," in Mohan Matthen, ed., *Oxford Handbook of the Philosophy of Perception* (New York: Oxford, in preparation).

Whatever perceptual constancy is exactly, Burge plainly means to put it to work. Perceptual constancies "are at the heart of the psychology of perception and the roots of objective reference" (233–34). Perception "requires perceptual constancies" (399). "A perceptual system achieves objectification by—and I am inclined to believe *only by*—exercising *perceptual constancies*" (408). Perceptual constancies are "the primary mark of perceptual objectification" (409). They are "necessary as well as sufficient for perceptual objectification and perceptual representation" (413). Let us look at how the argument goes.

I. STRAWSON AND EVANS

Burge distinguishes between two projects that concerned Strawson and Evans:

(1) "[E]xplaining constitutive conditions for having a *conception* of mind-independent entities *as* mind-independent," and
(2) "[*E*]*xplaining minimal constitutive conditions on objective representation* of the physical environment." (156)

Strawson and Evans are said to have "slid" from one to the other. Having the "conception" of objectivity seems to be a relatively sophisticated, reflective matter. We can assume self-consciousness on the part of someone who has this "conception," and more: we "can assume, uncontroversially, that we have conceptions of causation, of error, of a comprehensive spatio-temporal framework, of linguistic capacities and structures, and so on" (157). On the other hand, the project of establishing the constitutive conditions on objective thought cannot merely assume that any of these ideas have to be grasped by the subject. Particularly if one thinks that animals that altogether lack conceptual thought may be capable of objective representation, one will question whether any of these ideas have to be grasped in order for a subject to be capable of objective representation. Burge's proposal is that a causal theory of content, supplemented by the appeal to perceptual constancy, can explain how objective representation is possible. Perceptual constancy means that there is in some sense sameness through variation in proximal input. This means that in perception, "Specification of mind-independent and constitutively non-perspectival physical entities is separated out from the individual's sensory registration—the functioning state that encodes proximal sensory information" (398). In a suitable environmental context, "a systematic distinction, carried out in modular operations within perceptual systems, between what concerns an individual's receptors and what concerns a receptor-independent reality constitutes a type of objectification" (398).

---

[2] Anya C. Hurlbert, and Yazhu Ling, "If It's a Banana, It Must Be Yellow: The Role of Memory Colors in Color Constancy," *Journal of Vision*, v, 8, 787 (Sep. 23, 2005). See also, for example, Maria Olkkonen, Thorsten Hansen, and Karl R. Gegenfurtner, "Color Appearance of Familiar Objects: Effects of Object Shape, Texture, and Illumination Changes," *Journal of Vision*, VIII, 5, 13 (May 26, 2008): 1–16.

We do not need the kinds of material Strawson and Evans appealed to; perceptual representation, independent of thought, is already, determinately, fully objective. Any appeal to a purely causal theory of representation does face the familiar "disjunction problem": any perceptual system can be regarded as causally sensitive to many more or less proximal variables, so how can the system be said to be representing one thing rather than another? But, Burge says: "the causal structures of perceptual constancies suffice by themselves to solve most aspects of the Disjunction Problem. Only some of the most philosophically contrived 'disjuncts' are ruled out only by the role of zoological and ethological relations to the environment" (321n45).

Burge emphasizes the place of action in perceptual psychology, talking about "the deep role of perception in guiding animal life. Since perception guides action, it is not surprising that perceptual kinds mesh with animal activities" (324). Recall Gibson's view that "what we perceive when we look at objects are their affordances, not their qualities."[3] He said: "The affordances of the environment are what it *offers* the animal, what it *provides or furnishes*, either for good or ill."[4] Affordances are dispositions of things; they relate to what would happen were one to act thus-and-so. As Mark defined them, affordances are: "the functional utility of certain environmental objects or complexes taken with reference to individuals and their action capabilities."[5] They are relations between the environment and the observer. As Gibson puts it:

> An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is equally a fact of the environment

---

[3] James J. Gibson, *The Ecological Approach to Visual Perception* (Hillsdale, NJ: Lawrence Erlbaum, 1986), p. 134

[4] *Ibid.*, p. 127.

[5] Leonard S. Mark, "Eyeheight-Scaled Information about Affordances: A Study of Sitting and Stair Climbing," *Journal of Experimental Psychology. Human Perception and Performance*, XIII, 3 (August 1987): 361–70.

> and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer.[6]

Suppose we have a creature that is capable of representing only Gibsonian affordances. This creature does not have the capacity to represent the world as it is in itself, independently of its relations to the creature and its mental states and capacities for agency. Nevertheless, it may exhibit all the perceptual constancies that you or I exhibit. As Gibson says, visually detecting affordances is a matter of detecting "high-order optical invariants."[7] And the whole point about these high-order invariants is that the subject detects them through changes in proximal stimulation. The affordances provided by a chair, for example, remain invariant through changes in one's own position with respect to it, and vision detects these high-order invariants. Merely exhibiting perceptual constancy in perceptual representation is not enough for a grasp of the objective, intrinsic, categorical properties of the objects around one.

Even if you think of vision in computational terms, you could think that the point of the computations is to generate representations of affordances that the animal itself can use in action. The notion of an affordance specifies the object of perception: it does not depend on Gibson's idea that representation and computation are not aspects of perception. Burge argues that the computation of affordances must depend on detection of the qualities of objects:

> A reason to take functional attributives like *edible* to be psychologically parasitic on attribution of the physical properties is that similar principles for discerning bodies, size, motion, color, and so on govern visual systems among different species....By contrast, representations as of functional attributes, such as edibility or danger, are likely to apply to instances of food or danger in more species-bound ways. (323–24)

You could acknowledge this point, but argue that in ordinary vision, all that is represented at the level of the whole organism is the affordances of one's environment. That is all that the perception-action link demands. Of course, there may be perceptual subsystems, shared by many different species of animal, that use the qualities of objects as tools in generating these representations of affordances for the use of the organism. Those subsystems may themselves exhibit constancies in, for example, color or shape detection. But that is not enough to establish that color or shape is being represented at the level of the whole organism.

---

[6] Gibson, *op. cit.*, p. 129.
[7] *Ibid.*, p. 141.

Actually, in the case of many animals—take jumping spiders, for example—it is quite difficult to believe that they represent anything other than the affordances in their surroundings. The interests of most animals are severely practical. There is more to the world than the opportunities for action it provides to the jumping spider, but there is no reason to suppose that the jumping spider represents any of that "more."

If we ask what makes us confident that humans can perceptually represent anything other than the affordances that the world provides, one natural answer is that we have the conception of ourselves as occupants of a mind-independent reality. We take ourselves to have perception of a world which is as it is independently of us, and the affordances the world provides us are the consequences of a pair of factors: the way the world is in itself, and the ways in which we are able to act. What makes it the case that our perception is representing a world independent of us is not something that can be found simply by looking at the causal antecedents of perception: we have to consider the top-down impact of our conception of ourselves as occupants of an independent reality on the contents of individual perceptions. Strawson would have taken it absolutely for granted that there is no saying what it would be for a subject to have objective representations, unless we appeal to the subject's having the conception of itself as in a mind-independent world. The idea is that whatever we appeal to as a determinant of content, if we do not appeal to the subject's own conception of the mind-independent world, we will not be able to explain what *makes* the subject's perceptual representations into representations of a mind-independent world. If we have only the fact of a causal connection between the perception and the environment, we do not have enough to constitute representation of an objective reality. All that the subject is representing may be various opportunities for action that the world provides. This point is explicit in Evans. For example, in a passage actually quoted by Burge, he says that "the network of input-output connections which underlie the idea of an egocentric space could never be regarded as supporting a way of representing space (even egocentric space) if it could not be brought by the subject into coincidence with some such larger spatial representation of the world as is constituted by a cognitive map."[6] What is going on here is not a *slide*, but an argument that all that can make it the case that one has representations of an independent world is that those representations are being used in the context of a conception of oneself as in an independent world.

---

[6] Gareth Evans, *The Varieties of Reference* (New York: Oxford, 1982), p 163.

II. SEEING AND VISION SCIENCE

Burge's central point about perception is that his opponents uniformly went wrong in their neglect of perceptual psychology, and that we can derive a positive account of perception from perceptual psychology. How do we do this? Consider the terms that we use in common-sense talk about perception, such as "seeing that things are thus-and-so," or "seeing object *x*." Can they be *defined* in terms of the variables that perceptual psychology uses? In particular, can our ordinary notion of *seeing* be defined in terms of the notion of a perceptual representation? Of course, we are talking here about empirically grounded definition, rather than an analysis of what we ordinarily mean by terms like "seeing." Grasping Burge's answer to this question is central to understanding his positive picture of the way in which vision science can illuminate the phenomenon of seeing. On the face of it, Burge has a very straightforward reading of the relation between common-sense talk about seeing and vision science. There has to be more to seeing than merely having a perceptual representation of the type identified by vision science. The ordinary notion of seeing has a factive aspect: if you see that things are thus-and-so, they are thus-and-so. If you see Sally, then Sally must be there. Merely having the representation does not ensure that. Burge comments:

> although seeing is in a sense a natural kind, it is a hybrid kind. It is a psychological state that, *in each instance*, depends for being a seeing on entities and causal relations beyond the psychology of the individual. (389–90)

If this is right, then we might expect there to be a definition of seeing in terms of possession of a visual representation, together with a specification of the "entities and causal relations beyond the psychology of the individual" that are required for us to have a case of seeing. Burge does not attempt this, and there is a well-known difficulty facing the project. We will have to specify the causal relations required for seeing. Not just any causal relation will do, and there is no obvious way of specifying the right kind of causal relation. The situation is familiar from the case of propositional knowledge. Someone who has learned some cognitive science might argue as follows. The scientific study of cognition proceeds by postulating computational processing of representations. These representations can be as they are whether or not things are so in the environment. Therefore, *knowing* must be a "hybrid" state, consisting of possession of such a representation together with some other conditions, such as things being so, and a causal connection between that fact and the representation. Yet it has not proven possible to give the definition. Though Burge talks about *seeing* as a hybrid state, it may equally be impossible to specify the components of the alleged hybrid and how they are supposed to be put together.

The factive nature of our ordinary conception of *seeing* is one thing that pulls it away from the notion of a visual representation. On the other hand, vision science has a constitutive interest in the functioning of the brain that has no echo in common-sense psychology. This is not to say that common-sense psychology makes no implicit appeal to brain processes. On the contrary, our conception of singular causation for mental events is arguably shot through with the idea of a *process* connecting one mental event to another; and our common-sense picture of the "process" here arguably does need to be filled out by something like brain processes. But vision science has a constitutive interest in mapping the visual system in the brain, and common-sense psychology has no such constitutive interest in brain mapping. This is another way in which the variables of common-sense psychology pull apart from the variables of vision science. Burge's analysis of perceptual psychology does not give much weight to its constitutive interest in the functioning of the brain. He writes:

> The science is a psychological science, not a brain science—though it leaves open whether all psychology just describes brain states in psychological terms. Some of the brain mechanisms that underlie the psychological transformations described by the science—especially transformations that occur early in perceptual processing—are being discovered. But very little is known about the basis in the brain for even some of the most central types of perception—perception of color, location, shape, and so on. (369n3)

Does vision science "leave open" the existence of a level of implementation in the brain? Burge acknowledges Marr's "three levels" of explanation in vision science: (a) computation, (b) algorithm, and (c) implementation. He remarks, "The mainstream work in vision science that I discuss is primarily at levels (*a*) and (*b*)" (93n43). But it would be a mistake to suppose that a theoretical analysis of vision science can focus only on levels (a) and (b). Marr put the general emphasis of his approach like this: "I have laid particular stress on the level of computational theory, not because I regard it as inherently more important than the other two levels—the real power of the approach lies in the *integration* of all three levels of attack—but because it is a level of explanation that has not previously been recognized and acted upon" (my emphasis).[9] Vision scientists, even if in their published work they are not explicitly discussing physiological implementations, characteristically work with some sense of how in principle the computational systems they describe could be implemented in the brain. Ultimately, there has to be some account of how those boxes and arrows find a physiological realization. The very notion of *representation* that classical vision scientists work with has to have a brain implementation, however transient or scattered that implementation might be. But that is already enough to separate the variables of perceptual psychology from the factive variables of common-sense psychology, which has no such constitutive interest in mapping the brain.

---

[9] David Marr, *Vision* (San Francisco: W. H. Freeman, 1982), p. 330.

Computational processing occurs in many subsystems of perception, but Burge makes the point that perceptual representations, of the type studied by vision science, are used in the guidance of actions by the whole organism, and contribute constitutively to the organism's "perspective" on the world (370–72). Recognizing the constitutive role of brain mapping in vision science of course does not undermine this point, though Burge sometimes seems concerned that it will (189–90, 369). *Seeing*, as identified by common-sense psychology, is also a state that relates to the whole organism, rather than to some subsystem of the organism. The factivity of *seeing* on the one hand, and, on the other hand, the place in brain mapping of the notion of a *visual representation*, mean that we are here dealing with different phenomena. The situation seems to be that we have two variables relating to vision at the level of the whole organism, both legitimate, though one cannot be defined in terms of the other. Of course, the vision science illuminates the phenomenon of ordinary seeing. But the first thing is to recognize the difference.

Although perception has traditionally been thought of as involving experience of the world, Burge says, "I drop the association with consciousness" (368). This is, I think, right for the conception of visual representation used in vision science. It is not a persuasive analysis of the notion of *seeing* that we use in common-sense psychology. The underlying reason for this is that there is a constitutive role for *introspection* and *imagination* in our ordinary understanding of *seeing*. Our ordinary understanding of what it is to see something depends on our ordinary experience of seeing, and our ability to imagine other people seeing. Our ordinary conception of seeing is the conception of a subjective state.

Suppose you are using our ordinary concept of *seeing*, relying on introspection and imagination for your understanding of what we are talking about. What is different about different visual experiences one might have? To describe the qualitative character of a particular experience, what you do is describe the qualitative character of the objects you are seeing. When we understand *seeing* subjectively, we suppose that we describe the qualitative character of the experience by describing the qualitative character of the objects we see. The factivity of seeing is constitutive of it, on our ordinary subjective understanding of the phenomenon.

All this matters for the project of understanding objective thought because, if you are asking how objective thought is grounded in perception, it is, arguably, the common-sense conception of (for example) *seeing* that you need to appeal to, rather than the nonsubjective notion characterized by vision science. The content of the experience of seeing calls out to be characterized in terms of the objective, mind-independent, categorical properties of objects. The contents of visual representations, not constitutively related to experience, could just as well be characterized entirely in terms of affordances. Why do we take it that perception can ground objective thought, and not merely the representation of affordances? In characterizing the qualitative character of the conscious experience of seeing, we have to appeal to the qualities of the objects we see: their shapes and sizes and movements and so on. When we think of *seeing*, it seems perfectly obvious that this could ground objective thought about an independent environment, rather than merely a collection of affordances. Visual representation of affordances, on the other hand, could not of itself ground much more than thought about affordances.

This point is complemented by reflection on the distinction between ventral and dorsal streams in the visual system in the brain. The ventral stream is thought to sustain visual experience of the qualitative character of one's surroundings.[10] The dorsal stream, in contrast, has been conjectured to be dedicated to the detecting of affordances.[11] On this analysis, consider a blindsight patient whose ventral system is not receiving significant input relating to the left half of the visual field, while the dorsal system continues to receive some significant input from that half of the visual field. This patient is getting perceptual information about the affordances in the left half of the visual field. But that kind of process cannot explain how we form the conception of an independent world. For that, we also need the functioning of the ventral stream, sustaining conscious experience—ordinary seeing—of the things and properties around us. Burge writes as though the presence of perceptual constancies in the blind field is enough to establish that blindsight is the same kind of thing as ordinary seeing (374). But the differences between a representation of affordances in the dorsal stream and the factive, subjective state of ordinary seeing are at least as striking as the similarities, and they may be basic to an account of how objective thought depends on perception.

---

[10] A. David Milner and Melvyn A. Goodale, *The Visual Brain in Action* (New York: Oxford, 1995).

[11] Joel Norman, "Two Visual Systems and Two Theories of Perception: An Attempt to Reconcile the Constructivist and Ecological Approaches," *Behavioral and Brain Sciences*, xxv, 1 (2002): 73–144. See also Stephen E. Palmer, *Vision Science: Photons to Phenomenology* (Cambridge: MIT, 1999), p. 412.

III. DAVIDSON (AND EVANS)

Davidson said: "Our sense of objectivity is the consequence of [a] sort of triangulation, one that requires two creatures. Each interacts with an object, but what gives each the concept of the way things are objectively is the base line formed between the creatures by language."[12] This describes triangulation as a matter that itself depends on conceptual thought. Davidson later characterized the most primitive type of triangulation as "a prelinguistic, precognitive situation," involving "two or more creatures simultaneously in interaction with each other and the world they share."[13] In either formulation, why would you think that triangulation between two observers matters for objective representation? Here is one way to put the point. Suppose we have an animal that is not capable of triangulation with another observer on various aspects of the environment. The trouble is that such an animal might be capable of representing only affordances in the environment that are, as it were, affordances-for-it. By saying, "affordances-for-it," I do not mean that the animal is anywhere near exhibiting ordinary self-consciousness; we might be talking about, for example, a jumping spider. The point is that we have a perfectly well-defined notion of the threats or opportunities that the environment affords *that particular animal*. And the animal may be representing only those "affordances-for-it." It may not be capable of representing another animal as an agent, or representing any aspect of the environment as providing affordances for another animal. The animal might be capable of perceiving affordances, exhibiting perceptual constancy with respect to them, but not have any way of perceiving or cognizing the affordances available to other animals. Such an animal meets Burge's conditions on perceptual representation (368, 396–403). There is, nonetheless, a certain solipsism implicit in the representational system of such an animal. The point about Davidson's condition of triangulation is that it explains how an animal could exercise a capacity to represent its environment in a way that goes beyond that solipsistic perspective. An animal that interacts with others in its engagement with the environment may thereby show that it recognizes the environment as consisting of more than merely affordances-for-it. Now, Burge entirely rejects Davidson's appeal to triangulation:

---

[12] Donald Davidson, "Rational Animals," in *Subjective, Intersubjective, Objective* (New York: Oxford, 2001), p. 105.

[13] Davidson, "The Emergence of Thought," in *ibid.*, p. 128.

> The role that Davidson gives to triangulation in linguistic interpretation is filled much earlier in the ontogeny and phylogeny of the mental. Relevant triangulation occurs already in perception. Triangulation operates at two levels of abstraction. (272)

The "two levels of abstraction" that Burge has in mind are: (a) characterizations of the content of perception have to be integrated with explanations of the animal's needs and actions, and in that sense there is "triangulation among the explanatory objectives of biological explanations of whole-animal activity" and "fulfillment of whole-animal needs" (273), and (b) "a kind of triangulation that occurs in sensory perceptual systems. Such triangulation is pre-representational, pre-perceptual. It underlies perceptual constancies" (274). The idea here is that the object is "triangulated" from different proximal stimulations it generates.

As we have seen, this second level of "triangulation" cannot establish that the animal is representing anything other than affordances-for-it. But it may be that his first level of triangulation is meant to deal with just this point. On Burge's causal theory of perception, there will have to be an account of the biologically relevant foodstuffs and other animals and so on in the animal's environment. A full account of what is going on when the environment is filling the animal's needs and being acted on by the animal cannot appeal only to affordances, as if affordances could somehow exist unsupported in the perceiver's environment. The full scientific account must appeal to the actual things and so on that are there, providing these affordances. And the existence of this objective ground for the affordances is a constitutive condition on the animal's being able to represent at all. So the full scientific account of the constitutive conditions of the animal's being able to represent its surroundings must appeal to the objective things around it—the foodstuffs and other animals and so on—and not merely to the affordances-for-it provided by the environment. Isn't that enough to establish that the animal is representing objectively, rather than just solipsistically representing affordances-for-it? So we do not need Davidson's appeal to triangulation with another organism. This argument undoubtedly has a certain appeal, and seems to be what is motivating Burge here. Let us pause over its key move. This is the move from:

(1) the presence of the foodstuffs and other animals and so on (not merely the affordances they sustain) is a constitutive condition on the animal's being able to represent its surroundings, to

(2) the animal must therefore be representing those foodstuffs and other animals and so on (not merely the affordances they sustain).

Recall Burge's characterization of Individual Representationalism: "an individual cannot empirically and objectively represent an ordinary macro-physical subject matter unless the individual has resources that can represent some constitutive conditions for such representations" (13). The transition from (1) to (2) is simply an example of the slide Burge says is characteristic of Individual Representationalism: from "*X* is a constitutive condition on the capacity to represent objectively," to "*X* must be represented by anyone capable of objective representation." Once we resist this easy slide, there is nothing standing in the way of our saying that an animal could exhibit both of Burge's levels of triangulation without ever being capable of representing anything more than the solipsistic affordances-for-it. One natural way to explain why it seems so obvious that humans represent objectively, and are not confined to this solipsistic perspective, is that humans are capable of triangulation in Davidson's sense.

In the same chapter, Burge discusses Quine's various arguments for the conclusion that an ability to refer to concrete objects requires some grasp of criteria of identity for those objects. Naturally, Burge presses hard the idea that reference to objects is already present in perceptual representation, and that an account of the grounding of reference in perception must appeal not to patterns of proximal stimulation, but to the perceptual constancies that are invariant across patterns of proximal stimulation. Burge also extends his line of argument a bit, to argue that reference to objects as one and the same across quite long time periods need not depend on the subject's grasp of criteria of identity for objects. He gives a couple of examples of animals being able to reidentify particular bodies over periods of many months. He describes the famous Clayton et al. experiments in which scrub jays recovered cached food packets, and says: "The Clayton experiments show a capacity to track bodies over periods of 100 hours" (252n99).[14] And he presses the same idea in connection with dogs:

> A question is whether the dog's perceptual memory preserves the singular element in the perceptual representation. If the dog perceptually tracks the particular over time, or uses a remembered map in finding the bone, in addition to a capacity to respond to a type in the same way—matters that can be tested—the dog can reidentify a buried particular. (258)

---

[14] Nicola S. Clayton and Anthony Dickinson, "Episodic-Like Memory during Cache Recovery by Scrub Jays," *Nature*, CCCXCV (Sep. 17, 1998): 272–74.

Suppose a scrub jay caches a food packet and an experimenter carefully replaces it with a qualitatively indistinguishable, yet numerically distinct, food packet. Burge's view is that there is a significant sense in which the scrub jay has been fooled when it returns to the cache. Suppose a dog buries a bone and later returns to dig it up, only to find that it has rotted to the point where even a dog would find it inedible. Burge's view is that there is a significant sense in which the dog has got it right. A contrasting view is that animals have no sentiment about the identity of particular food packets or particular bones. The scrub jay and the dog would not share Burge's assessment of the situation. All that matters for the animal is the affordance. There is no reason why evolution should have fitted out the animals with a capacity to represent particulars that goes beyond their needs.

I think this is an example of a curious and persistent tendency in Burge to suppose that absolutely all cases in which (a) the reliability of a cognitive mechanism depends on its being sensitive to how things are with particular objects, must be cases in which (b) the representations generated by the mechanism make reference to particular objects. Another example occurs in Burge's discussion of Evans, where Burge breezily translates a remark of Evans about the conditions for the accuracy of a *mechanism* for generating existential contents (namely, on Evans's view of it, the mechanism constituted by the visual system) into a remark about the conditions for the accuracy of the *contents* generated by the mechanism (185).[15] Since Evans takes the accuracy of the mechanism to depend on its sensitivity to how things are with the particular objects seen by the observer, Burge supposes Evans to be committed to supposing that vision generates singular representations relating to those objects, rather than merely existential representations. But, of course, that does not follow at all. Similarly, an animal might have a representational system that represents only the types of affordances to be found at particular locations, and the reliability of the mechanism generating those representations might depend on its sensitivity to how things are with particular objects. But it does not follow that the representations generated by the mechanism are themselves making reference to those objects.

IV. INDIVIDUAL REPRESENTATIONALISM

Let us look again at Burge's way of stating the idea that he uses to frame the book: Individual Representationalism. According to Burge, the core of Individual Representationalism is that "an individual cannot empirically and objectively represent an ordinary macro-physical subject matter unless the individual has resources that can represent some constitutive conditions for such representation" (13). As I mentioned earlier, there is a generality in this formulation that makes it useful for marshaling together a wide variety of apparently distinct views. After all, a wide range of positions is available on what the "constitutive conditions" are on objective representation. And given any one position on what the constitutive conditions are on objective representation, any one variety of Individual Representationalism can select a different one of those constitutive conditions as being what the subject must represent in order to represent objectively. This makes the general formulation very useful for pulling together a bewildering diversity of philosophical positions—virtually every position held in twentieth-century philosophy—as one or another version of Individual Representationalism (111–283).

---

[15] Compare Evans, *op. cit.*, p 125.

What about Burge's own position? Is it a form of Individual Representationalism? According to Burge, in some central cases the constitutive conditions on an ability to perceptually represent *X's* include this one: (*) the presence of *X's* in one's environment (85). So if all the constitutive conditions are met, then, of course, one will have the ability to represent (*) the presence of *X's* in one's environment. So on Burge's view, representing *X's* will require that one can represent one of the constitutive conditions for such representations: namely, the presence of *X's*.

This is not a superficial difficulty of formulation. As we have seen, Burge's line of thought depends on the move from "environmental conditions *X* are constitutively required for perceptual representation," to "the subject must be representing environmental conditions *X*." That is what allows Burge to think that perceptual representation across animal species must be representing more than affordances. Moreover, stated in this abstract way, the principle that "objective representation requires the ability to represent some of the constitutive conditions for representation" seems quite unmotivated. So Burge's formulation of the issues in terms of Individual Representationalism goes quite deep into the way in which the book tries to shift the burden of proof onto its opponents. If, say, you are drawn to a causal theory of representation, then you will accept the principle of Individual Representationalism, for the reason I just gave. But it will not be the abstract principle that drives your acceptance of the causal theory of representation. It is rather your acceptance of the causal theory of representation that drives your acceptance of the abstract principle. Burge writes as though his opponents have unthinkingly accepted the abstract principle and then, in its thrall, have cast about to find ways of implementing it. This seems to me to get things round the wrong way, for his opponents as well as for Burge. In the case of Quine, for example, it is not that he is a fanatical believer in the abstract idea that objective representation requires an ability to represent the conditions of the capacity for objective represention. What drives Quine is, rather, the attempt to understand the normativity of meaning in purely physicalist terms. That itself may be a motivation open to challenge, but it is a different motivation.

To put the point another way, as we have just seen, it would be possible to be an Individual Representationalist while holding, with Burge, that the objectivity of propositional thought is entirely grounded in the objectivity of perceptual representation. This means that it is not right to frame the main issue of the book in terms of the doctrine of Individual Representationalism, because the issue between Burge and his opponents would still be there. The principal issue is whether the objectivity of propositional thought is wholly grounded in the objectivity of perceptual representation.

The general problem with the book's positive account of objectivity is that Burge overestimates the explanatory power of his own approach to representation: a causal theory supplemented by an appeal to perceptual constancies. To see the general problem, notice that his conditions on objective representation are amply met by the jumping spider (genus *Portia*), as Burge acknowledges (515–16). To put the point bluntly, it seems to me fantastical to suppose that the jumping spider is representing a mind-independent reality in anything like the senses that Strawson or Evans or Davidson or Quine had in mind when they talked about "objectivity." What the jumping spider cares about is dinner, mating, and finding the way home. It represents the opportunities and threats its environment offers it, and there is no point or purpose in its representing anything else. These opportunities and threats are really there in the world, but they are defined only in relation to the needs of the spider. They are not spider-independent aspects of the world. The spider does not represent a spider-independent world. Humans (and perhaps other animals), in contrast, are capable of representing a species-independent reality. The ongoing problem is to understand how this is achieved, and, most basically, what this ability consists in. Appeals to capacities for spatiotemporal thought, grasp of criteria of identity, the role of conscious experience, and joint attention or triangulation are all serious contributions. They are unlikely to be simply displaced by an appeal to perceptual constancy, though that too is an important idea, and highlighting it is a valuable contribution of *Origins of Objectivity*.

What seems to me excellent about the book, and likely to be of lasting value, is that it brings into clearer view these different approaches to the explanation of objective content and the question how much work can be done autonomously by a "bottom-up" causal theory of perceptual content.

JOHN CAMPBELL

University of California, Berkeley

UNITED STATES POSTAL SERVICE

Statement of Ownership, Management, and Circulation

(All Periodicals Publications Except Requester Publications)

1. Publication Title: The Journal of Philosophy

2. Publication Number: 0 0 2 2 - 3 6 2 X

3. Filing Date: 9/26/2011

4. Issue Frequency: Monthly

5. Number of Issues Published Annually:

6. Annual Subscription Price: Individuals $45.00; Institutions $[illegible]

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4®)

The Journal of Philosophy, Columbia Univ
1150 Amsterdam Ave, Box [illegible]
New York, NY 10027

Contact Person: Pam Ward

Telephone (Include area code): 212 [illegible] 4619

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer)

Same

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address)

The Journal of Philosophy Inc.
1150 Amsterdam Ave
New York, NY 10027

Editor (Name and complete mailing address)

Akeel Bilgrami, President
The Journal of Philosophy
1150 Amsterdam Ave New York, NY 10027

Managing Editor (Name and complete mailing address)

[illegible], Managing Editor
The Journal of Philosophy 1150 Amsterdam Ave
New York, NY 10027

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

Full Name	Complete Mailing Address
The Journal of Philosophy	Columbia University
	1150 Amsterdam Ave
	Mc 4972
	New York, NY 10027

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☐ None

Full Name	Complete Mailing Address

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, September 2007 (Page 1 of 3 (Instructions Page 3)) PSN 7530-01-000-9931 PRIVACY NOTICE: See our privacy policy on www.usps.com

---

13. Publication Title: The Journal of Philosophy

14. Issue Date for Circulation Data Below: May 2011

15. Extent and Nature of Circulation			Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)			3500	3500
b. Paid Circulation (By Mail and Outside the Mail)	(1)	Mailed Outside-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	700	674
	(2)	Mailed In-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	2400	2247
	(3)	Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®		
	(4)	Paid Distribution by Other Classes of Mail Through the USPS (e.g. First-Class Mail®)		
c. Total Paid Distribution (Sum of 15b (1), (2), (3), and (4))			3100	2921
d. Free or Nominal Rate Distribution (By Mail and Outside the Mail)	(1)	Free or Nominal Rate Outside-County Copies included on PS Form 3541		
	(2)	Free or Nominal Rate In-County Copies Included on PS Form 3541		
	(3)	Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g. First-Class Mail)		
	(4)	Free or Nominal Rate Distribution Outside the Mail (Carriers or other means)		
e. Total Free or Nominal Rate Distribution (Sum of 15d (1), (2), (3) and (4))			150	150
f. Total Distribution (Sum of 15c and 15e) ▶			3250	3071
g. Copies not Distributed (See Instructions to Publishers #4 (page #3)) ▶			250	429
h. Total (Sum of 15f and g) ▶			3500	3500
i. Percent Paid (15c divided by 15f times 100) ▶			95 45	95 12

16. Publication of Statement of Ownership
☐ If the publication is a general publication, publication of this statement is required. Will be printed in the 5/2011 issue of this publication.
☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner

Pamela C. Ward

Date: 9/15/11

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

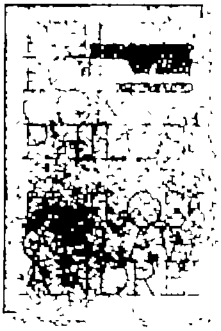
PS Form 3526, September 2007 (Page 2 of 3)

Truth

Alexis G. Burgess & John P. Burgess

"Surveying an important and much-discussed field, and balancing the nature and logic-and-paradox sides of truth in an easy-to-grasp fashion, this book is a major achievement. It is easily the best first step into truth studies available, but also an excellent short review for experts. It is difficult to write clearly and convey serious ideas about truth to nonspecialists, but the authors have somehow pulled it off."
—JC Beall, University of Connecticut

Cloth $22.95 978-0-691-14401-6

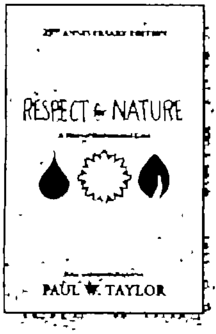


Philosophy of Law

Andrei Marmor

"This superb book, written by one of the world's foremost philosophers of law, is a highly welcome addition to the jurisprudential literature, one marked by its immense pedagogical value and scholarly excellence. It provides, in clear, accessible language, an informed, sympathetic treatment of all the main issues and theories within the discipline of contemporary analytic legal philosophy"
—W. J. Waluchow, author of *A Common Law Theory of Judicial Review*

Cloth $22.95 978-0-691-14167-1



Respect for Nature

A Theory of Environmental Ethics

25th Anniversary Edition

Paul W. Taylor

With a new foreword by Dale Jamieson

"When it first appeared, *Respect for Nature* was at once recognized for the important, groundbreaking work it was. It was deservedly a major influence in the then newly developing field of environmental ethics. Time has only confirmed this first opinion and it is good to have the book back before us."
—Michael Ruse, editor of *Philosophy after Darwin*

Paper $24.95 978-0-691-15024-6

CAMBRIDGE

NEW *and* NOTEWORTHY TITLES *in* PHILOSOPHY

New in Paperback!

Leibniz
An Intellectual Biography
Maria Rosa Antognazza
$29 99 Pb 978-1-107-62761-1 664 pp

Kant, Religion, and Politics
James DiCenso
$95 00 Hb 978-1-107-00933-2 280 pp

Philosophy and Religion
Volume 68
Anthony O'Hear
$32 99 Pb 978-1-107-61598-4 320 pp

Aristotle's *Nicomachean Ethics*
A Critical Guide
Jon Miller
$85 00 Hb 978-0-521-51448-4 296 pp

New in Paperback!

Plato's Cosmology and its Ethical Dimensions
Gabriela Roxana Carone
$32 99 Pb 978-1-107-65704-5 332 pp

Agro-Technology
A Philosophical Introduction
R. Paul Thompson
$90 00 Hb 978-0-521-11797-5 258 pp
$28 99 Pb 978-0-521-13375-3

Critical Pragmatics
An Inquiry into Reference and Communication
Kepa Korta and ***John Perry***
$95 00 Hb 978-0-521-76497-1 192 pp
$29 99 Pb 978-0-521-74867-4

Proof Analysis
A Contribution to Hilbert's Last Problem
Sara Negri and ***Jan von Plato***
$90 00 Hb 978-1-107-00895-3 272 pp

Logical Dynamics of Information and Interaction
Johan van Benthem
$95 00 Hb 978-0-521-76579-4 384 pp

Political Philosophy in the Twentieth Century
Authors and Arguments
Catherine Zuckert
$95 00 Hb 978-1-107-00622-5 368 pp
$29.99 Pb 978-0-521-18506-6

The Cambridge Companion to Nozick's *Anarchy, State, and Utopia*
Ralf M. Bader and ***John Meadowcroft***
$90 00 Hb 978-0-521-19776-2 328 pp
$29 99 Pb 978-0-521-12002-9

New in Paperback!

The Cambridge History of Eighteenth-Century Philosophy
2 Volume Boxed Set
Knud Haakonssen
$90 00 2 Pb books 978-0-521-28088-4 1,424 pp

Tocqueville: The Ancien Régime and the French Revolution
Jon Elster and ***Arthur Goldhammer***
$75 00 Hb 978-0-521-88980-3 320 pp
$19 99 Pb 978-0-521-71891-2

Prices subject to change

www.cambridge.org/us

CAMBRIDGE
UNIVERSITY PRESS

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 6/7
JUNE/JULY 2011

page

The Dewey Lectures 2007: *Self and Self-Understanding* Tyler Burge

287 Lecture I: *Some Origins of Self*
316 Lecture II: *Self and Constitutive Norms*
339 Lecture III: *Self-Understanding*

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

---

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER. Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2011 by the Journal of Philosophy, Inc. ISSN 0022-362X

UNIVERSITY OF CALCUTTA ALIPORE CAMPUS

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 6/7, JUNE/JULY 2011

SELF AND SELF-UNDERSTANDING[*]

LECTURE I: SOME ORIGINS OF SELF

I will reflect on constitutive features of selves—especially a certain sort of self-understanding. This self-understanding is the main topic of these lectures.

I

'Self' is a technical term, refined from ordinary usage. Ordinary usage is, however, very close to what I want. A definition from the *Oxford English Dictionary* runs, "Self: a person's essential being that distinguishes the person from others, especially considered as the object of introspection or reflexive action; a person's particular nature." Kant characterized a *person* as "what is conscious of its numerical identity, of its self, in different times."[1] This conception of *person* is very

---

[*] This work is a revision, with some expansion, of the Dewey Lectures, given at Columbia University, December 2007. I am grateful to Christopher Peacocke for valuable criticisms in Spring 2011 of the last section of Lecture I and all of Lecture III; and to Denis Bühler for saving me from an error and prompting an argument in Lecture II.

[1] Immanuel Kant, *Critique of Pure Reason*, A361; see also *Metaphysics Mrongovius* 29:911 in Karl Ameriks and Steve Naragon, eds., *Lectures on Metaphysics* (New York: Cambridge, 1997), p. 276. Kant's formulation in the *Critique* is not ideally specific and might qualify as ambiguous. It is clear from context that Kant means the consciousness to include consciousness at a given time of the self as it is at different times. It is also clear that Kant intends the consciousness to be noninferential, and in my terms *de re*. Kant shows less interest in the *diachronic* implications of his formulation than one might hope. He relies on these implications in the Third Paralogism to argue that cognition of self over time is necessarily empirical and cannot meet requirements of rationalist theories of self. In my "Memory and Persons," *The Philosophical Review*, CXII, 3 (July 2003): 289–337, note 50, I sketch why I reject Kant's argument that self-cognition of self over time is necessarily empirical. In notes from his lectures, some characterizations of *person* or *personality* omit reference to trans-temporal self-consciousness altogether, citing only a being conscious of its identity in different states. See *Metaphysics* L_1 28:276–77; *Metaphysics Dohna* 28:680. Both passages are in Ameriks and Naragon, eds., *op. cit.*, pp. 87–88, 381. See also *Critique of Practical Reason* 5:87, in Mary J. Gregor, ed., *Practical Philosophy* (New York:

0022-362X/11/0806-07/287–383

close to the notion of a *self* that interests me. It is a purely psychological notion that specifies a self-consciousness with diachronic reach. Regardless of the ontology of selves—whether or not they are purely psychological beings—the *concept* self is a psychological concept.

Kant's concept contrasts with Strawson's concept person. Strawson's concept is close to the common-sense notion of person—roughly, the sort of bodily being—paradigmatically human being—that normally matures into a self-conscious critical reasoner with moral capabilities. Strawson's *concept* is *not* purely a psychological concept. It *entails* that persons have physical as well as psychological characteristics. These two notions—the OED/Kant notion and Strawson's notion—typify a distinction that I draw between selves and persons.

Selves are *loci*, indeed agents, of psychological activities, and agents that can engage in, and are subject to, certain valuations—the distinctive valuations associated with persons. Selves evaluate themselves and other selves for rationality, critical rationality, morality, social cooperation, character, creativity, grace. Being an agent and topic of such valuations is part of what makes them selves. Since selves constitute the particular psychological natures of persons, selves' being agents and topics of such valuations is part of what makes persons *persons*.

In engaging in these types of valuation, selves become the subject matter of self-knowledge and self-understanding. So the valuational and cognitive powers of selves are intertwined. Some of the types of evaluation that I listed are constitutive to being a self. Selves and persons are constitutively capable of evaluating, and engaging in, critical reason. Arguably, they are constitutively capable of being moral. A certain reflexive cognition is constitutive to these evaluations. Critical reason and morality constitutively depend on self-understanding. My topic is the psychology and epistemology of the kind of self-understanding that is required for critical reason and morality, and that is itself constitutive to being a self, and a person.

Cambridge, 1996), p. 210. Kant thought that self-consciousness necessarily involves inner sense. There is ample evidence that he assumed that inner sense necessarily has past and future as well as present applications. So these failures to mention trans-temporal capacities do not reflect a different position from the one in *Critique of Pure Reason*. But they do show that he does little to develop this aspect of the notion. In *The Metaphysics of Morals* 6.223—in Gregor, *op. cit.*, p. 378—Kant distinguishes moral personality—a free being with reason—from psychological personality, explained in terms of consciousness of self in different states. (See also *Metaphysics* L_1 28.276–77, *op. cit.*) The two notions are connected, however, inasmuch as a free being with reason must be capable of remembering and anticipating acts if acts are to be imputed to the individual's moral personality.

First, some introductory points. I am interested in *constitutive* matters—in the *nature* of selves—those features of selves that make them selves.[2] I characterized selves as certain *loci* of *psychological* realities. The fact that self is a purely psychological notion does not entail, or even much encourage, the view that selves lack physical natures. The psychological nature of the *concept* cannot legislate these ontological matters. I will, however, be focusing on the *psychological* nature of selves. I bracket any further, nonpsychological aspects of their nature.

It is natural and right to connect psychological discussion with discussion of the brain. I will have almost nothing to say about the brain. I think that a psychological framework is necessary to finding anything in the brain that is of psychological interest. Ultimately, psychology and neuroscience are collaborative enterprises. Each must provide checks and balances on the other. But a psychological framework will provide us with more than enough material to work with here. The issues about understanding and value that I will be discussing *must* be connected to psychological inquiry. We cannot understand these issues in nonpsychological terms. Perhaps this situation will change. But I doubt it.

The tradition of discussing selves is complex. There are several closely related concepts that figure in the tradition—soul, spirit, mind, subject, conscious subject, self-conscious subject, rational animal, rational being, critically rational being, ego, person. Several of these will appear in my discussion. First, I will return to contrasting my notion of self with that of person.

Strawson's elaboration of the notion of person has strongly influenced philosophy in the last half-century. I think that this influence is deserved. But it has led to misconceptions about relations between psychological and physical attributions. I will not have time for detailed discussion of Strawson's views. I will just describe them and state some attitudes toward them.

Strawson developed his concept of persons to combat dualist conceptions of mind.[3] According to his concept, a person essentially has both psychological and corporeal attributes. Strawson's key claim is that use of the concept person, with its essential application to a corporeal being, is necessary for understanding psychological concepts. In other words, application of psychological concepts would be incoherent and unintelligible if they were not applied to something, a person, that has corporeal as well as psychological attributes.

---

[2] I discuss natures in my *Foundations of Mind. Philosophical Essays, Volume II* (New York: Oxford, 2007), pp 1–3 and passim; and in *Origins of Objectivity* (New York: Oxford, 2010), pp. 57–67 and passim. For working through passims, see the indexes

[3] P F. Strawson, *Individuals* (1959) (London: Routledge, 2002), chapter 3.

I believe that none of Strawson's arguments, or any by his followers, justifies this claim.[4] All depend on views about knowledge or individuation of content that are neither plausible nor well supported. I do think that Strawson's notion person is useful. What I reject is a set of arguments for holding that all psychological notions are conceptually dependent on a prior notion of a corporeal being, a person. I will take psychological notions on their own terms. I emphasize that these are conceptual, constitutive points. I will not be defending an ontological position.

It is only because mature persons naturally have certain psychological capacities that we evaluate them for critical rationality, take them to have a special moral status, and so on. Not all persons have these states and capacities at all times. Six-month-olds are persons. But if a being naturally and constitutively did not have or develop psychological states and capacities beyond those *exhibited* by six-month-old human beings, they would not be persons.[5]

Perhaps persons *develop* into having or being selves. They would do so by developing the competencies that meet certain conditions for falling under norms constitutive of being selves, such as norms for engaging in critical reasoning or for being morally responsible. Then persons would not have or be selves during all times when they, the persons, exist. Demented human individuals that lack psychological capacities beyond chimps are persons, but only because their natural capacities have been distorted by age, damage, or disease. We understand our concept person partly by reference to certain natural psychological competencies. The reference is to naturally having the competencies in a mature state, not to having them throughout a person's existence. I think of selves as beings that actually have the competencies that make persons the valuable beings that they are. Thus selves are types that set standards for being a person. The standards allow for prolepsis and retrospection.

---

[4] Strawson argues for his views in *ibid.* See also his *The Bounds of Sense* (1966) (London: Routlege, 2006), p. 163ff For followers, see Gareth Evans, *The Varieties of Reference* (New York: Oxford, 1982), pp. 208ff, 237ff.; John McDowell, *Mind and World* (Cambridge: Harvard, 1994), p. 100ff.; John Campbell, *Past, Space, and Self* (Cambridge: MIT, 1995), p. 92ff; Quassim Cassam, *Self and World* (New York: Oxford, 1997), passim; and most of the essays in José L. Bermúdez, Anothony Marcel, and Naomi Eilan, eds., *The Body and the Self* (Cambridge: MIT, 1995).

[5] I write of *naturally* having competencies, or of *natural* competencies, to allow that a self or person could be damaged or diseased so that performance that would realize the competencies is blocked. If a self permanently loses such competencies, it no longer exists or the person no longer is a self.

It is compatible with the main part of my discussion to hold that mature selves are just phases in the lives of persons. There are intuitions that pull in that direction. On the other hand, I find it natural to reify—to say that someone has gone out of existence when a person still exists, but has irrevocably lost those psychological capacities that made the person a person. I am inclined to count selves as individuals. I think that natural instincts and even explanatory and normative considerations suggest being liberal in allowing existence to different types of being, even where they overlap in seemingly messy ways. Here, these issues will not matter. I focus on constitution, nature, essence. I shall speak of selves, for the sake of exposition, as individuals.

Like most psychological notions, the notion of self bears complex constitutive relations to a wider physical reality. As far as I know, selves may necessarily depend on a physical body. I think, however, that there is no evident *conceptual* or even apriori necessity that selves have physical properties, although all selves that we know of depend in some way on their bodies.

Selves are necessarily conscious, at least some of the time. A type of consciousness presupposed by all other types is phenomenal consciousness.[6] Phenomenal consciousness consists in there being some way that it is like to be in a mental state. Phenomenal consciousness is an occurrent condition. It is the psychological bedrock of selfhood. An individual's stream of consciousness is interrupted, in sleep for example. But phenomenally conscious psychological life is where the acts, cognitions, and evaluations that most matter to us occur.[7]

Underlying conscious life is, of course, a vast system of unconscious psychological states and activities. In empirical explanation, unconscious elements are primary. But unconscious psychological states and occurrences are constitutively relevant to selves only insofar as they bear certain relations to conscious states and occurrences.

I will concentrate here on the *representational* nature of selves. Selves are subjects with representational competencies, states, events, and acts. They are subjects with perceptual systems, capacities for inference, beliefs, intentions, perceptions, occurrent thoughts, decisions.

---

[6] The notion of phenomenological consciousness is vividly illustrated in Thomas Nagel, "What Is It Like to Be a Bat?" *The Philosophical Review*, LXXXIII, 4 (October 1974): 435–50.

[7] Compare William James, *The Principles of Psychology*, vol. 1 (Cambridge: Harvard, 1981 [1890]), chapter 10, pp. 284–85, which identifies the self with ".. either the entire stream of our personal consciousness, or the present 'segment' or 'section' of that stream, according as whether we take a broader or a narrower view." He calls the narrower, more present segment the "innermost centre within the circle, or sanctuary within the citadel, constituted by the subjective life as a whole. Compared with this element of the stream, the other parts, even of the subjective life, seem transient external possessions, of which each in turn can be disowned, whilst that which disowns them remains....It is the home of interest....It is the source of effort and attention, and the place from which appear to emanate the fiats of the will."

Now I introduce a technical term. The *point of view* of an individual is the system of representational states and occurrences that are, in a certain sense, correctly imputable to the individual.[8] Nonhuman animals as well as persons and selves have points of view. Imputability is a partly intuitive, partly theoretical notion. I start with intuitive points. *Nonpsychological* entities are not part of a point of view. Irretrievably unconscious modular *psychological* states and processes are also not part of a point of view. They are not imputable to the individual, though they occur in the individual's psychological subsystems. Potentially or actually conscious perceptions, memories, beliefs, intentions, emotions, decisions, imaginings are part of a point of view. So are attitudes in the Freudian unconscious.

I believe that the key to the nature of selves lies in reflecting on powers distinctive to their points of view. These powers include competence for a certain type of reflexive self-attribution of elements in the point of view. Such self-attributions give the type of point of view that is distinctive of selves a *multi-tiered* structure. I shall explore this structure. It is the structure of self-consciousness and representation of self.

II

I begin by discussing some background conditions and some representational antecedents of the representation of self that is distinctive of selves and persons.

Every living system can differentiate itself from other things. It expels foreign matter as waste, ingests other bodies and not its own, protects itself from external threats, and so on. These capacities are not in themselves either representational or psychological. They form the functional background for self-interest and representation of self.

Perception is the lower border of sensory representation.[9] Not all sensory systems are perceptual. Hence not all are representational.

---

[8] By extension, I count the representational contents and nonrepresentational conscious features of the relevant psychological states the individual's point of view. I emphasize that '*point of view*' and '*perspective*', which I use interchangeably, are *not* to be construed as indicating anything essentially perceptual or empirical. One can have a purely mathematical point of view or perspective. Points of view and perspectives are ways of representing (thinking about, perceiving, remembering) a subject matter. I confine these terms to "ways," and states marked by those ways, that are imputable to an individual, rather than subsystems of individuals.

[9] I defend this claim and the associated conception of representation in *Origins of Objectivity*, especially chapters 8 and 9. A state is representational in my sense if and only if specification of the state in terms of veridicality conditions could ground serious, correct explanation.

Sensings of heat or light by amoebae or worms and the sensors in muscle tissue that signal stress are not representational in my sense. The explanatory paradigms that drive the theory of perception differ fundamentally from those that drive explanations of the sensitivities in these sensory systems. Explanation of perception makes essential reference to veridicality conditions. Explanation centers on how the system represents the surrounding environment accurately, to the extent that it does. Of course, even the simplest sensory systems exhibit functional success and failure. A sensory system is functionally successful when its response to stimulus is associated in the environment with the macro-qualities that it evolved to respond to. But in nonperceptual sensory systems, success is only functional good fortune. There is no explanatory value in invoking veridicality conditions.

By contrast, perceptual systems are explained in terms of the formation of states that are about the distal environment and that are veridical or not. The mark of a perceptual system is *perceptual constancy*. Perceptual constancies are abilities systematically to represent given distal features, by sensory means, *as* those same features, even though the proximal stimulations caused by those features vary radically. For example, most perceptual visual systems can represent something as having a given size as it moves closer or farther away—thus as it causes radically different proximal stimulations. Perceptual constancies occur in the visual, auditory, and touch systems of a wide variety of animals, including the visual systems of certain insects.

As noted, all living organisms show self-interest, broadly understood. I am interested in antecedents of *representation* of self. The most primitive representational antecedent is constitutive to all perceptual and actional representational content. This antecedent is an *egocentric index*.[10] In their most primitive forms, such indexes are not conceptual. Applications of egocentric indexes anchor all perception and all primitive actional states. In primitive forms, they are not used or applied *by the individual*. They are, however, a primitive type of ego-related representation. These primitive forms become incorporated into propositional perceptual beliefs and propositional intentions. Then they are conceptual.

---

[10] Another term for such representation is '*de se* index' For discussion of such indexes, see "Memory and Persons," and *Origins of Objectivity*, pp 199–201, 287–88, and passim. For other discussions, see Marc Jeannerod, *The Cognitive Neuroscience of Action* (Cambridge: Blackwell, 1997); Jeannerod, "To Act or Not to Act: Perspectives on the Representation of Actions," *Quarterly Journal of Experimental Psychology*, LIIA, 1 (1999): 1–29; J. D. Crawford, W. P. Medendorp, and J. J Marotta, "Spatial Transformations for Eye–Hand Coordination," *Journal of Neurophysiology*, XCII (2004). 10–19.

In perception, egocentric indexes refer to the individual—or, more commonly, some position on the individual or some time at which the perception occurs—as an *origin* with respect to which representation of other entities is mapped. Egocentric indexes single out an individual, position, or time in a context-dependent way. For example, an index might mark a spatial origin in relation to which other entities are spatially related in a perception. A nest-perception might represent the nest as to the left at such and such a distance from the origin of the perception. The origin is perhaps a position between the eyes of an insect. Or a perceptual memory of a food source might represent it as having been visited at some time interval from the time represented by an egocentric index as the present time of the memory representation.

Like the self-concept, egocentric indexes are immune to reference failure. They automatically mark an origin of reference that occurs whenever a state with the relevant index is instantiated. Unlike applications of the first-person concept, occurrent applications of egocentric indexes are normally not acts. Applications depend purely on their role in a larger system of representation. Unlike applications of the first-person concept, not all egocentric indexes *refer* to an individual or ego. Many refer only to a position or a time.

Egocentric indexes have two constitutive functions. One is the function just discussed—to index an origin for a framework of representation. The other is to type-individuate an aspect of psychological states that treats the origin as having immediate ego-relevance. All representational psychologies function to privilege the individual's needs, goals, and perspective, and function to provide the representational basis for serving the individual's needs, goals, and perspective—doing things for the individual from the individual's own motivations or other powers. The applications that realize the first of the two constitutive functions of egocentric indexes—applications to spatial or temporal positions—are always themselves privileged in an individual's psychology. They function as privileged in that they function in the psychology to have psychologically immediate ego-relevance. They mark spatial origins of perception and spatial positions from which the individual's actions are initiated. They mark positions that the individual must protect in serving the individual's own needs, goals, and perspective. They mark times, present times, for initiating action and calibrating memories and anticipations. They mark times that are privileged with respect to serving the individual's needs, goals, and perspective.

Like all psychological representational content, the ego-relevant content of spatial and temporal egocentric indexes marks representational functions that are in turn grounded in representational competencies. The ego-relevant content marks competencies to treat indexically anchored positions as privileged in realizing functions to distinguish and serve the individual's own needs, goals, and perspective. All representational psychologies have these functions and competencies. All mark temporal or spatial positions indexically in an ego-relevant way. The competencies that underlie these primitive ego-related representations are enriched in individuals that have complex psychologies, until they are part of the suite of competencies that are constitutive of persons and selves. Egocentric indexes are the ur-antecedents of concepts of self.

Of course, an insect cannot spell out this ego-relevance. It has no attributives that apply to psychological states or content. Egocentric indexes, however, are part of the representational content of—and help individuate—psychological states that systematically relate a framework origin to the individual's needs or perspective, noninferentially and as immediately as the system allows. This functional relation to the individual's needs, goals, or perspective is what makes an index—and the psychological state that it helps individuate—egocentric.

The relation to the individual's needs, goals, or perspective usually runs through capacities mediating perception and action. For example, the psychological system might be such that if an object is perceived to be approaching the origin, the animal will take protective measures. Or certain feeding actions might be initiated if a memory marks a food source as having been visited at a certain temporal interval from the present.

Egocentric indexes and their two functions are constitutive to every representational psychology. Every representational psychology contains some states that include them. Powers to realize these functions help mark an individual with a representational mind as an individual.

Fulfilling these functions involves a representational capacity to distinguish and privilege the individual's own needs, goals, and perspective. By distinguishing and privileging the individual's needs, goals, and perspective, egocentric indexes provide a representational basis for serving them. The unity of an individual's representational psychology consists partly in having representational powers that mark off and unify that psychology by fulfilling the individual's needs and goals, and otherwise serving its representational perspective.

As noted, egocentric indexes are ancestors of the first-person concept, expressed by the word 'I', that represents persons and selves. The indexes' infallible reference, their anchoring a representational framework, and their direct ego-relevance all presage aspects of the first-person concept. Having egocentric indexes does not require a capacity to *represent* anything *as* having psychological states. In this respect, they differ from a full concept of self.

III

In the remainder of this lecture, I discuss four capacities that mark milestones in the developmental prehistory of distinctively *self*-representation. The first three are more complex exercises of egocentric indexes. The exercises are psychological couplings of egocentric indexes between and within sensory modalities. These couplings are, I think, early precursors of a certain capacity for objectification that is characteristic of selves and persons. I have in mind the capacity to think of oneself from first- and third-person points of view. In their most primitive forms, these couplings occur among sensory capacities. They are nonconceptual. Nonhuman animals exhibit at least the first capacity, and possibly all three. They are thus nodes in the "animal" roots of selves. The fourth capacity is a type of memory that yields another kind of objectification constitutive to selves and persons.

Let us consider the first of these four capacities. In the last three decades, ethology and developmental psychology have been astir over beginnings of self-awareness. Some of this discussion is onto important junctures in the prehistory of personhood and selfhood. The literature has, however, persistently blurred important psychological distinctions. Much of it vastly overstates the directness of connection between the phenomena being theorized about and self-awareness in any sense that involves representation of a representer's own psychology.

I start with what is called '*the mirror test*'. A mark is applied to a bodily surface without an individual's being aware of the application. The mark is placed so that the individual can see it only in a mirror. An individual passes the mirror test if after a short time, it acts as if its own behavior is the source of the behavior shown in the mirror.[11] For example, the individual might touch the mark. Chimpanzees, elephants, and dolphins pass the test. Only a few gorillas in captivity have passed it. It was long thought that although monkeys can learn to use mirrors to spot other objects, they cannot pass the test. Recently, there is some evidence that rhesus macaques pass it. Human children first pass it between 16 and 24 months.

---

[11] Gordon G. Gallup, Jr., "Chimpanzees: Self-Recognition," *Science*, CLXVII, 3914 (Jan. 2, 1970): 341–43; Gallup, "Self Recognition in Primates: A Comparative Approach to the Bidirectional Properties of Consciousness," *American Psychologist*, XXXII, 5 (May 1977). 329–38, Sue T. Parker, Robert W. Mitchell, and Maria L. Boccia, eds., *Self-Awareness in Animals and Humans: Developmental Perspectives* (New York: Cambridge, 1994); Diana Reiss and Lori Marino, "Mirror Self-Recognition in the Bottlenose Dolphin. A Case of Cognitive Convergence," *Proceedings of the National Academy of Sciences*, XCVIII, 10 (May 8, 2001): 5937–42.

Passing the test has been taken as evidence for self-awareness.[12] The relevant notions of *self* and *awareness* need clarification. There is no clear sign that in passing the test an animal represents itself as having psychological states. The experiments do not indicate possession of what I count a self-concept or first-person concept. Still, there are steps in those directions.

Passing the mirror test rests on bidirectional couplings in the individual's psychology between visual perceptions and representations in kinesthetic proprioception. The individual can match what it sees with what it senses in its own movements. The individual's psychology translates between visual representations and kinesthetic/proprioceptive representations as of the same movements—tokens and types. The translation is systematic. The mapping seems to use a continuing image of the whole body, built through proprioception in the individual's learning history. Mirrors are not essential for such couplings, but passing the mirror test evinces them.

The systematic character of the couplings and the fact that they operate on a whole-body image differentiate the behavior exhibited in passing the mirror test from the behavior of feeling an irritant on a body part, finding it visually, and brushing it off. This latter type of behavior also involves couplings between different sense modalities. It does not, however, suggest visual identification of something, codified in the body schema, as the whole individual.[13] Visually locating and dealing with an irritant does not depend on systematic whole-body coordination between vision and proprioception. Visually identifying as such what has been mapped proprioceptively *as of a whole individual* is a kind of reflexive "self" identification. Here the "self" is the whole individual. The individual is identified in terms of bodily characteristics. The identification is, as far as the experiments show, at the level of intermodal connections between perceptual capacities, not necessarily effected through propositional thought.

The connection between vision and *kinesthetic* representation makes possible not only the individual's representation of its whole body, but makes possible a kind of representation of the body *as the individual's own*—that is, something like an *attribution* of *ego-indexing* to the individual's own bodily actions. Let me explain this point step by step.

---

[12] Gallup, "Self-Recognition: Research Strategies and Experimental Design," in Parker, Mitchell, and Boccia, eds., *op. cit.* Gallup, the originator of the test, gives the following experimental condition on exhibition of self-awareness: "An organism is self-aware to the extent that it can be shown capable of becoming the object of his own attention."

[13] For scientific discussion of the notion of body schema, see J. Paillard, "Body Schema and Body Image: A Double Dissociation in Deafferented Patients," in Gantcho N. Gantchev, Shigemi Mori, and Jean Massion, eds., *Motor Control. Today and Tomorrow* (Sofia, Bulgaria: Academic Publishing House, 1999), pp. 197–214. For philosophical discussion, see Brian O'Shaughnessy, "Proprioception and the Body Image," in Bermúdez, Marcel, and Eilan, eds., *op. cit.*

First, ordinary visual representational content is egocentrically indexed. But such content *attributes* to *seen* particulars only physical or functional characteristics (shape, position, color, danger, food, and so on). Primate vision represents body parts and bodily movements as such. It probably also differentiates purposive from nonpurposive movements.[14]

Second, the bodily movement seen in the mirror causes not only visual representation but a kinesthetic sense of the movement. The representational content of kinesthetic perception, like that of visual perception, is egocentrically indexed—marking the origin of the perception's perspective. But there is an important difference between proprioception and vision. In proprioception, egocentric indexes also tag the *perceived* entities. Proprioception marks the perceived bodily movement as having the same relevance to the individual's needs and perspective as the proprioceptive perceptions themselves have. A function of proprioceptive perception is to track the animal's movements *as* the animal's own. So whereas, by the point of the previous paragraph, ordinary visual representations are singly indexed, kinesthetic perceptual representations are always *doubly* indexed. In addition to the index, or indexes, for the perspectival origin of the perception, kinesthetic perceptual content associates an egocentric index with the *perceived* particulars. Proprioceptively perceived bodily movements are marked as privileged with respect to the individual's needs, goals, and perspective.

Third, the proprioceptively perceived body parts and movements are mapped onto a standing proprioceptive body schema. Like all proprioception, this schema is doubly indexed: both representational origin and proprioceptively perceived entities are indexed.

Fourth, for individuals who pass the mirror test, mirror experiences set up an intra-psychological coupling between visual representations of bodily movements and kinesthetic action representations of the same movements. These couplings induce a further coupling between the visual representations of body parts and positions on the body schema.

---

[14] B. Hare et al., "Chimpanzees Know What Conspecifics Do and Do Not See," *Animal Behaviour*, LIX, 4 (April 2000): 771–86, Laurie R. Santos, Jonathan I. Flombaum, and Webb Phillips, "The Evolution of Human Mindreading: How Nonhuman Primates Can Inform Social Cognitive Neuroscience," in Steven M. Platek, Julian Paul Keenan, and Todd K. Shackelford, eds., *Evolutionary Cognitive Neuroscience* (Cambridge: MIT, 2007), pp. 433–56. A lot of theory in this domain over-attributes representation of psychological states to apes and very young children, when attributing representation of teleological states suffices to account for the evidence.

Fifth, these couplings provide a causal pattern that grounds new representational content. There is a new association of egocentric indexes with *objects* of *visual* perception. Thus *seen* objects—say, the movement of a body part—are directly and systematically coordinated in the psychology with the individual's needs, goals, and perspective. The movement is seen as ego-privileged. Single-indexed and double-indexed visual perceptions represent the same movement. But the latter are constitutively linked to kinesthetic and body-image representational contents.[15]

The application in vision of single-indexed and double-indexed representational contents to the same movement is not a propositional identity. It is a coordination between perceptual representations in the psychology. But the visual system functions to treat the different representations as representing the same movement.[16]

Passing the mirror test marks three developmental milestones. They may have been achieved before the test is passed. Passing the test just shows that they have been achieved.

First, by associating an egocentric index with a purportedly visually perceived physical entity, the psychology invests that entity with the same ego-significance that the origin of the visual perception has. Recall that the second function of egocentric indexes is to mark a direct psychological relation between the individual's needs, goals, or perspective and what they index. What is new is that this indexing is associated with a *visually* perceived entity. No more in this new step than in the most primitive egocentric indexing does the individual use psychological attributives that might figure in a *theory* of mind. No psychological states are represented (referred to or indicated) by egocentric indexes.

---

[15] The association of single-indexed and double-indexed visual representations is informative. Such associations can in principle be mistaken. The usual visual representation of the form

[ego$_i$]that$_1$movement

comes to be associated with a visual representation of the form

[ego$_i$]that$_1$[ego$_i$]movement.

Both represent the same movement. At the level of a psychology that passes the mirror test, the egocentric indexes may well refer to the whole body rather than merely a body part.

Earlier I said that egocentric indexes are infallibly referentially successful in their primary function. In associating egocentric indexes with perceived entities, a perceptual system is fallible inasmuch as perceptual representation of perceived entities can fail. So association of an egocentric index with a purportedly perceived entity can participate in a failure to connect with purportedly perceived reality. However, even such an index marks a relation in the psychology of an origin in perception to the individual's needs, goals, and perspective. Occurrences of an egocentric index cannot fail in representing the origin or in connecting a purported perceptual attribution to the individual's needs, goals, and perspective.

[16] This association ramifies the perceptual constancies that mark all perception. I conjecture that a visual system cannot produce doubly indexed representational contents on its own. A doubly indexed visual representation always depends on some intra-psychological connection to a naturally doubly indexed representation in another perceptual modality—specifically, proprioception—or in a conceptual application. In proprioception, the individual perceives his or her own bodily arrangements or movements, and naturally perceives them as being ego-related.

Second, the individual's psychology associates occurrences of its egocentric indexes with the individual's whole, visually perceived body, not just with an indexed anchor *position*. As noted, the proprioceptive body schema constitutively marks a perceptually represented whole body, and does so in a doubly indexed way. It is not part of the *nature* of the visual system to doubly index anything. In coming to associate an egocentric index with a visually perceived body, the psychological system has an intra-psychological functional connection between the anchor of the visual perceptual representational content and the perceived body. (Of course, perceived body parts can also be associated with an egocentric index and mapped as parts of the doubly indexed whole body.) As with the first milestone, what is new lies in the linking of visual competence with double-indexing that is already present in another system.

A third step is implicit in the second, and is, I think, the most significant one. Passing the mirror test evinces a step toward *objectification* of self in two respects. One is intermodal. The development of connections among kinesthetic proprioception, the proprioceptive body schema, and visual perception allows the perceiving individual to associate egocentric indexes with the same individual as *perceived from the perspectives of different perceptual modalities*. The perceived individual is perceived as having ego-significance, in this limited way, from different perspectives; and the perspectives are linked with one another and with actional representations.

The other aspect of objectification concerns visual representation. Couplings among vision, kinesthetic sense, and body-schema representation make possible an informative connection of single-indexed visual perceptions with double-indexed visual perceptions of the same entities. A visual perception of a face—anchored as usual by an egocentric index—may represent a face as just another face. Then when the intermodal connections are established, the singly indexed perception comes to be informatively associated with a visual perception that attributes an egocentric index to the same face. So the second perception is double-indexed. The face is seen from two perspectives. (Proprioception is always double-indexed and does not provide two perspectives.) One perspective characterizes one's face, but not as ego-related. The other characterizes the identically appearing face as the individual's own. The case is a primitive, preconceptual ancestor of "Aha, that's me!" It is a primitive ancestor of representing oneself from both first-person and third-person perspectives.

To describe these advances as an emergence of *self*-consciousness or *self*-representation would be misleading. Egocentric indexes are associated with an object of representation. But no psychological states are attributed to it. So no individual is represented *as* a self. Reflexivity in representation should not be confused with genuine *self*-consciousness or *self*-representation. Still, the reflexivity and coupling of different points of view shown in the mirror test is a *step* toward genuine self-representation.

IV

The psychology evinced in the mirror test is a simple case of intermodal and intramodal perceptual couplings of egocentric indexes. Such couplings occur in more complex forms. Each form is of interest in itself. Each adds a distinct type of objectification in egocentric indexing. I will discuss two other couplings—*imitation* and *joint attention*.

Imitation starts from the beginning of human life.[17] Children imitate the expressions of parents from the first few minutes after birth. Purposive social copying plays a major role in the emergence of more refined systematic couplings among an individual's sensory modalities and, of course, in the development of language. *Copying* is matching in which one specific type of behavior in one individual is caused by, and specifically explained by, the sensing—perceptual or not—of the same type of behavior in another individual. Copying divides into several interestingly different subspecies, of which imitation is the most important for the development of a self-concept. *Imitation* is active, purposeful, goal-directed copying of a specific form of behavior. Examples of nonimitative forms of copying are *mimicry*—an automatic sensory-motor type of copying—and *emulation*—purposive copying whose point is to match a behavioral *result*, not necessarily the behavior itself. It is disputed whether nonhuman animals ever imitate as opposed to emulate.[18] Nonhuman animals are certainly more oriented toward results than toward coordinating behaviors themselves. From 12 months onward, human children differ dramatically from other primates through greater orientation to imitation, and less to emulation.

---

[17] Andrew N. Meltzoff and M. Keith Moore, "Imitation of Facial and Manual Gestures by Human Neonates," *Science*, CXCVIII, 4312 (Oct. 7, 1977): 75–78; Alison Gopnik and Meltzoff, "Minds, Bodies, and Persons: Young Children's Understanding of the Self and Others as Reflected in Imitation and Theory of Mind Research," in Parker, Mitchell, and Boccia, eds., *op. cit.* This copying may be associated with mirror neurons—neurons activated both when an individual observes specific types of activity and when the individual performs the same specific types of activity. Mirror neurons were discovered by Vittorio Gallese et al., "Action Recognition in the Premotor Cortex," *Brain*, CXIX, 2 (April 1996): 593–609. Mirror neurons are popularly construed as instantiating "theory of mind" or "mind reading." This construal over-intellectualizes the phenomenon. The function of mirror neurons has not been agreed upon. It is natural to think that they are central to copying behavior. But their presence in some primates, macaque monkeys, has not been associated with copying

Imitation again involves systematic coupling among sensory modalities, and between sensory modalities and the representational actional system. Suppose that the imitating individual has a visual perception, $(V)\underline{[ego_1]that_1F}$. ('(V)' designates the visual modality. 'F' indicates an attributive mode of presentation that is perceptually applied to the token behavior *A* of the imitated individual. I assume that the egocentric index is spatial. I ignore the inevitable temporal index. The subscripts on 'ego' and 'that' mark occurrent applications of the index and the demonstrative-like capacity, respectively.) The individual adjusts his action—and his psychology adjusts representation of action—so that it is of the same *type* as the imitated action. $(V)\underline{[ego_1]that_1F}$, referring to *A*, is taken as the paradigm model for the individual's imitation. The imitating individual produces a type-similar imitating action A_I in such a way that it satisfies a representation $\underline{F}$ as of the same type of action as the imitated action *A*. So the psychology produces in its actional system and, presumably in its kinesthetic proprioceptive system, doubly indexed representations $(Act)\underline{[ego_1]that_2[ego_1]F}$ and $(K)\underline{[ego_1]that_2[ego_1]F}$. ('Act' indicates actional representational mode; 'K' stands for kinesthetic mode.) The actional representation anticipates the imitating act A_I. The kinesthetic representation is caused by A_I. Given that the imitation is successful, acts *A* and A_I are in fact of type F. Imitation is a purposive coupling between the agent's own act A_I and another individual's type-similar act *A*. The coupling of the visual representation $(V)\underline{[ego_1]that_1F}$, which perceptually represents the act *A* that is to be imitated, with representations $(Act)\underline{[ego_1]that_2[ego_1]F}$ and $(K)\underline{[ego_1]that_2[ego_1]F}$ constitutes a social analog of the objectification that emerged in the mirror test.

In imitation, the individual's psychology does not achieve the full reflexivity that it does in passing the mirror test. There is no coupling of a single-indexed visual perception with a doubly indexed actional or proprioceptive perception *of the same token act*. And there is no informative coupling within vision of an ordinary perception of a seen token act with a doubly indexed visual perception *of the same token act*.

---

[18] The distinction between imitation and emulation is due to Michael Tomasello, "Do Apes Ape?" in Cecilia M. Hayes and Bennett G. Galef, eds., *Social Learning in Animals: The Roots of Culture* (New York: Academic, 1996), pp. 319–46.

Only singly and doubly indexed representations of different tokens of the same type of action are coupled.

Although imitation begins earlier in human development, it requires more, in this respect, than passing the mirror test does. Through the mirror, the couplings of seen acts with proprioceptively sensed acts are *given* to individuals that are equipped to register them. Since the individual is tracking his or her own acts visually, what is represented visually is in lock step with what is proprioceptively represented. By contrast, in imitation the individual must adjust his actions to a sequence of actions that he or she does not control. So the type matching must be made through adjustment in imitation. It is not there simply to be apprehended.

A little reflection indicates how fundamental dyadic copying, including imitation, is in infant life. Parent and child enter into games of ritualized turn taking. The range of cases in which coupling between the child's visual representations and its proprioceptive representations, for achievement of sameness of representational type, must be enormous. These couplings are the backbone of early self-objectification.[19]

In human beings, dyadic copying occurs earlier than the triadic coordination commonly involved in *joint attention*. Joint attention is widely regarded as the culmination of coordinated social interaction in the pre-linguistic phases of human development.[20] Joint attention emerges in the middle of the second year of human life. Roughly, joint attention is a capacity to share attention with another, usually toward a third object, with some mutual awareness of the shared attention.

The intermodal couplings in joint attention are complex topics in themselves. But I will not discuss their form or content here. As with passing the mirror test and imitation, joint attention does not in itself involve attribution of psychological states.[21] What is attributed to others is essentially purposive behavior. The purposiveness need not be represented as deriving from psychological states. Both imitation and joint attention do, however, provide further elements in the developmental background for robust *self*-representation.

---

[19] Cf. Vasudevi Reddy, "Before the 'Third Element': Understanding Attention to Self," and Sue Leekam, "Why Do Children with Autism Have a Joint Attention Impairment?" both in Eilan et al., eds., *Joint Attention. Communication and Other Minds* (New York: Oxford, 2005), pp. 85–109, 205–59.

[20] Malinda Carpenter, Katherine Nagell, and Tomasello, "Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age," *Monographs of the Society for Research in Child Development*, serial no. 255, LXIII, 4 (1998).

[21] Although attending is a psychological act, and although in joint attention individuals represent acts (gazes) that are in fact guided by attention, the acts are first represented as purposive activity, not as psychologically guided purposive activity. See note 14 above.

We mature into a psychological conception of self. But we seem not to start there. Our initial ego-representations are bodily. They involve attribution of purposive activity. The activity is overt physical action. The *psychological* attributions young children come to make are, I believe, primarily built on perceptual representations of bodies, bodily agents, bodily attributes, supplemented with teleological representations of purposive or goal-directed acts. Teleology is not psychology. Our representation of ourselves as psychological beings grows from representation of ourselves and others as purposeful denizens of a physical world.

V

Intermodal perceptual coupling of egocentric indexes is probably the most primitive representational antecedent of self-representation and self-understanding. A seemingly more advanced antecedent, which I will discuss for the remainder of this lecture, is a certain type of memory. Selves are necessarily extended in time. I think that a point-event self is an incoherent notion. Acting and experiencing take time. A self-concept must be associated with representations of oneself as extended in time. A certain type of memory underlies such representations. This type is systematically associated with a certain type of anticipation.

I will rely on a rather extensive taxonomical background.[22] Please bear with me. The delineation of types of memory is, of course, a matter for empirical psychology. I will present a classification that is a conceptually clarified version of what can be found in empirical work.

I use the term 'memory' only for a representational capacity. I intend 'representational', as usual, in my relatively demanding sense—requiring an explanatory use for appeal to veridicality conditions in type-identifying states and capacities. I do not take retentions of the effects of classical conditioning, or so-called muscle memory, to count as memory, properly so called.

Iconic memory, working memory, and other types of short-term memory have short automatic decay times and, in the latter two cases, fairly strict load capacities. I shall be mainly concerned with a certain type of long-term memory. Long-term memory encodes relatively significant matters, usually tagged through attention. Long-term memory effectively retains representational content indefinitely, with no known automatic decay times. All these types of memory occur in a wide variety of animals, including many that lack propositional attitudes.

---

[22] This taxonomy improves on one I use at the outset of "Memory and Persons."

Let us distinguish between *experiential* memory and *nonexperiential* memory. Experiential memory is remembering a particular entity *x* in a direct *de re* way. The particular can be a mental event, a physical event or physical object, or a scene. If I see rain in Salisbury and thereby remember that it rained in Salisbury, without coding the memory perceptually, my memory is not experiential, though it derives from perceptual experience. If I think *cogito* and remember that I did, but do not remember my thinking it, the memory is not experiential.[23]

The division between experiential and nonexperiential memory cuts across most of the types already mentioned. Iconic memory is always experiential. But working memory, short-term memory, and long-term memory can be experiential or nonexperiential. Propositional thought can occur in any of these latter three types of memory. Experiential memory can be propositional or nonpropositional, as long as it is *de re* of a particular or particulars.

I will focus on a type of long-term experiential memory—*episodic memory*. *Episodic* memory is conscious, long-term, experiential memory whose perspective functions to pick out a past particular, representing it as it was at a specific past time.[24] Tulving characterized episodic memory picturesquely, as a type of mental time travel to the past.

Episodic memory evolved more recently than other memory systems. Its neural basis lies in the prefrontal cortex and other neocortical regions. It is more vulnerable to disease, injury, and aging than other types of memory. It seems to emerge late in child development, possibly not until 24 months, after imitation and joint attention emerge. Tulving conjectured that episodic memory is unique to humans. This conjecture is controversial. It is hard to test whether long-term animal memory of specific events is *conscious*. Something like episodic memory, not necessarily with consciousness, has been plausibly attributed to birds, dolphins, and primates.

There are two primary distinctions between generic long-term experiential memory and its subspecies, episodic memory. One is that the latter must be conscious, whereas the former need not be.[25] The other is that episodic memory must function to represent a particular *as it was at a specific past time*. Long-term experiential memory can be *de re* of a particular without picking out the particular episodically. A nonepisodic long-term perceptual experiential memory could have a perceptual residue that is too generic, or too much a perceptual composite of experiences, to single out the particular as it was at a given time.

---

[23] To be experiential in this sense, a memory need not have the format of *perceptual* experience. It can derive from intellectual events. I am using 'experiential' in a very broad sense.

[24] The past particular could be an object or scene, as well as an event. Discovery and investigation of episodic memory as a distinctive psychological phenomenon derives from the work of Endel Tulving, *Elements of Episodic Memory* (New York: Oxford, 1983); "Remembering and Knowing the Past," *American Scientist*, LXXVII, 4 (July–August 1989): 361–67; "Episodic Memory and Autonoesis: Uniquely Human?" in Herbert S. Terrace and Janet Metcalfe, eds., *The Missing Link in Cognition: Origins of Self-Reflective Consciousness* (New York: Oxford, 2005), pp. 3–56 Cf. also Alan D. Baddeley, Martin A. Conway, and John P. Aggleton, eds., *Episodic Memory New Directions in Research* (New York: Oxford, 2002).

Both generic long-term experiential memory and episodic memory are *de re*. *De re representation* is representation that picks out a particular on the basis of a noninferential, not purely attributional or descriptional representational competence.[26] Nonepisodic experiential memory can be *de re* recognition, without representing an entity as it was at any specific time. Episodic memory is not only *de re* reference to past particulars. The *de re* representation retains a past representational state that functioned to represent a particular as it was at a corresponding past time; and the *de re* representation is accurate if and only if it derives from, and accurately represents, the particular as it in fact was, and was represented, at the relevant past time.

The notion of episodic memory, with its evocative metaphor of mental time travel, can perhaps be made vivid by contrasting it with another type of long-term memory. Imagine an animal that buries food and returns to uncover it months later. Whether the unearthing behavior indicates episodic memory—or even nonepisodic experiential memory—is so far completely open, even laying aside the issue of consciousness. Consider the following updating mechanism, which is thought to occur in many nonepisodic types of memory. The animal perceives the burying. A nonpropositional representation with the content, that food in such and such a location now, is laid down. Memory updates this content as time passes: that [remembered] food in such and such a location now. The remembered food is expected to be currently in the location. Although memory derives from past experience, there is no memory of the burying, or of the food as it was at the time of the burying, or of the burial time itself.

---

[25] Long-term experiential memory is easier to investigate empirically insofar as one does not require consciousness. There is little agreement about how to determine whether animals' memories are conscious. Often experimenters stipulate that they are studying animal '*episodic-like*' memory precisely to avoid the issue of consciousness. Cf. Nicola S. Clayton and Anthony Dickinson, "Episodic-like Memory during Cache Recovery by Scrub Jays," *Nature*, CCCXCV (Sep 17, 1998): 272–74; Charles Menzel, "Progress in the Study of Chimpanzee Recall and Episodic Memory," and Bennett L. Schwartz, "Do Nonhuman Primates Have Episodic Memory?" in Terrace and Metcalfe, eds., *op. cit.*, pp. 188–241. Tulving's characterizations of episodic memory confine *representata* of such memory to events But memory of bodies and scenes can be included, as long as memory represents them as they were experienced at a particular time.

[26] For discussion of *de re* attitudes, see my "Belief *De Re*," this JOURNAL, LXXIV, 6 (June 1977): 338–62; reprinted with "Postscript to 'Belief *De Re*'," in *Foundations of Mind*; also "Five Theses on *De Re* States and Attitudes," in Joseph Almog and Paolo Leonardi, eds., *The Philosophy of David Kaplan* (New York: Oxford, 2009), pp. 246–316.

Such updating use of memory does not rule out having representational content in past-tense form: that food once buried in such and such location. This nonepisodic memory still lacks *de re* reference to the past. Human children younger than 24 months show no signs of episodic memory. Before that time, they can recount past events as being past. But their retellings are generic, script-like.[27] Events are situated within a script, but not on the basis of an individualized autobiographical experience. Human adults who have lost episodic memory through neural damage or aging can tell historical, even autobiographical narratives. But the narratives are like scripts learned from hearsay or reading. The narratives rest on memories that do not enable the individual to re-experience past events, or relive the past.

Lacking episodic memory, with its *de re* representation of particulars as they were in the past, an individual would be unable to represent such particulars in the way that the individual experienced them. The form and order of the individual's own past would be lost. For an individual with propositional attitudes, the loss would be a deficit in understanding. The individual's understanding of his or her representational past would be at best indirect.

Earlier I discussed objectification through couplings between different perceptual modalities. Such couplings coordinate representations of the same entity from the perspectives of different perceptual modalities. Such couplings provide a basis for weak ego-objectification, inasmuch as an egocentric index is attached to representation *associated* with a *perceived* entity.

---

[27] Katherine Nelson and Janice Gruendel, "Children's Scripts," in Nelson, ed., *Event Knowledge: Structure and Function in Development* (Hillsdale, NJ: Lawrence Erlbaum, 1986), pp. 21–46; Nelson, "Emerging Levels of Consciousness in Early Human Development," in Terrace and Metcalfe, eds., *op. cit.*, pp. 116–41, Teresa McCormack and Christoph Hoerl, "Memory and Temporal Perspective: The Role of Temporal Frameworks in Memory Development," *Developmental Review*, XIX, 1 (March 1999): 154–82 During ages three to five, there is a leap in richness, detail, and specificity in narration. Stories are made up as well as recounted. Temporal indexicals such as yesterday and tomorrow emerge in the third year, later than indexicals like I and you. I disagree with accounts of episodic memory, such as the last one listed, that maintain that episodic memory constitutively involves not only *de re* representation of past entities as they were at past times, but also self-consciousness and a capacity to 'interrelate arbitrary events'. Such requirements are held to be necessary to 'give meaning' to the *de re* elements in episodic memory. I believe such views commit the *philosophical* error of hyper-intellectualizing the phenomenon, along lines analogous to those that I criticize, with respect to perceptual reference, in *Origins of Objectivity*, chapter 6. The *de re* representation of past events as they were at specific past times requires a capacity to distinguish events in different instances of script-like or recurrent phase-like sequences This capacity requires either a capacity to localize one event in a sequence, or a capacity to distinguish differences in types of events within otherwise similar sequences. But it does not require a use of allocentric, decentered frameworks, or a capacity to refer to arbitrary past events—much less self-consciousness.

This phenomenon can be replayed in perceptual experiential memory. *Ordinary*, nonproprioceptive, perceptual experiential memory—episodic or not—is analogous to visual perception in being singly indexed. It does not in itself associate an index with any remembered entity. It just carries the egocentric index of the purported original perceptual experience, or some conceptual analog. Nothing in the remembered scene is represented as having privileged ego-status for the rememberer. There are, however, experiential memories that are the memory-cousins of doubly indexed visual perception that passes the mirror test. Such doubly indexed perceptual experiential memories are autobiographical.[28] I call an experiential memory that associates an egocentric index, pre-conceptual or conceptual, with the individual rememberer, represented as an entity in the remembered situation, an '*autobiographical experiential memory*'. An individual's autobiographical perceptual experiential memory not only functions to preserve a previous egocentric index functioning as perceptual origin. The memory also represents the rememberer, placing the rememberer in the remembered situation; and it marks the rememberer, thus represented in memory, as having ego-significance.

One way in which an autobiographical perceptual experiential memory can place the rememberer in the remembered scene is as an object of perception. An individual can remember and associate an egocentric index with him- or herself as seen in a mirror—the memory analog of passing the mirror test. But an autobiographical visual experiential memory need not place the rememberer in the remembered scene as an object of perception. One can remember oneself *as* experiencing something other than oneself. Retaining a visual memory is distinct from retaining a visual memory in which one is also remembered as the visual perceiver. Such memory probably requires *conceptual* capacities. Perhaps most autobiographical *visual* experiential memories among selves, and *mature* persons, will be of this sort.[29]

---

[28] Episodic memory need not be conceptualized. It can be purely perceptual. Episodic memory also need not be meta-representational. It need not *attribute* psychological states, beyond the "attribution" involved in associating egocentric indexes with remembered entities, in autobiographical cases. Even when autobiographical, it may make only physical and teleological attributions to entities, marking them as having ego-significance. A common mistake among psychologists who discuss episodic memory is to claim that episodic memory constitutes a meta-perspective on one's perspective. There is no basis in the phenomenon for claiming that it is representation *of* perspective. Episodic memory retains and enables one to call up perspective. The perspective is *reused, not referred to* in the simpler forms of episodic memory.

I call a memory that is veridical, and whose veridicality includes preservation of the representational perspective of a representational state that is retained in the memory, 'memory *from the inside*'. I am interested in autobiographical episodic memories from the inside.

In the case of *perceptual* episodic memories, this notion of being from the inside is fairly straightforward. The memory retains the angle of perception—the perceptual perspective—on the subject matter that the original perception had. Not all perceptual memories are from the inside, in this sense. One could have a memory, in perceptual format, as of some subject matter. The memory could get right intrinsic properties and relations of the perceived subject matter. But the memory could be from a perceptual perspective that one never had. The memory is partly veridical, partly not. One might remember the layout of the scene quite correctly, but from a perceptual perspective outside of one's actual past perspective.[30]

The notion of retaining the perspective of a memory that is nonperceptual—for example, retaining a perspective of a particular occurrent inference or occurrent thought—is similar, even if less phenomenologically vivid. Retaining the cognitive perspective of the original intellectual occurrence—and being an episodic memory from the inside—is veridically retaining the mode and representational content of that occurrence.[31] Inasmuch as the memory is episodic, it must be also be a *de re* memory of that occurrence.

---

[29] All experiential memories that place the rememberer in the remembered situation but do not do so through remembering the rememberer as an object of perception must have a conceptual, or at least post-perceptual, attributional element. Remembering oneself *as* a perceiver requires a representation as of perception. Perception itself cannot represent perception as such.

[30] A common but dramatic way in which a perceptual memory may be partly veridical, but not be from the perceptual perspective of the original experience, is for a memory to have a positional origin that is a rotation from that of the original experience. I have memories of an important childhood event from an impossible perspective twenty feet above the ground. Often a memory nonveridically places the rememberer as a perceived participant in the scene. In the psychological literature, memories from an external observer's position are called 'observer memories'. They are contrasted with 'field memories', those that take up the same positional perspective of the original experience. See Sigmund Freud, "Screen Memories" (1899), in James Strachey, ed. and trans., *The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume 3* (London. Hogarth, 1953–74); Georgia Nigro and Ulric Neisser, "Point of View in Personal Memories," *Cognitive Psychology*, xv, 4 (1983): 467–82; John A. Robinson and Karen L. Swanson, "Field and Observer Modes of Remembering," *Memory*, I, 3 (1993): 169–84.

[31] There may be more to the original perspective than its cognitive aspects—what Frege called 'coloring', for example. I focus on retaining cognitive aspects of the perspective.

Thus to be from the inside, an episodic memory must be veridical, and the veridicality must include retention of the perspective of the original perceptual or intellectual "experience."[22]

I have described autobiographical episodic memory from the inside because I believe that it constitutes another step toward self-objectification. It extends the process into the temporal dimension in ways that passing the mirror test, imitation, and joint attention do not.

There is dispute in the developmental and ethological literatures over how *meta-psychological representation*—representation of psychological kinds and instances as such—emerges. I think that the matter is not well understood, and will not discuss it here. Certainly by age three or four, human children have psychological concepts and apply them to themselves and others. I believe that when married to meta-psychological representation, autobiographical episodic memory from the inside is a momentous type of self-understanding that is constitutively necessary for being a self and for being subject to norms of morality and critical reason.

When autobiographical episodic memory attributes psychological states, it contributes to specifying one's own psychological history. It enables one to revive one's psychological past in some of the same form that it took. It connects one's present and past perspectives. Unlike updating memory, it provides a basis for attributing a contrast of perspectives. When an individual can *refer* to perspectives as such, episodic memory provides a direct, *de re* way for the individual to represent him- or herself as a psychological being with changing commitments.

VI

I conclude by reflecting on how the type of episodic memory that I have been discussing gives point to the Kantian conception of self that I cited at the outset. Recall that Kant wrote of a being that "is conscious of its numerical identity, of its self, in different times." What sort of diachronic consciousness is involved? Why does it matter? I will first make some remarks about consciousness. Then I discuss the relevant consciousness's significance.

Kant is right that consciousness figures prominently in our conception of selves. The relevant consciousness is representational and meta-psychological. The representation is both of the self and of its psychological states and events.[33] It attributes psychological states, and it associates an ego-like representation of them with them.

---

[22] This notion of veridically retaining the original perspective must allow for more or less. Most "veridical" memory involves some omission and some errors. Exact sameness with the contents of the perceptual point of view rarely or never occurs.

Consciousness is a distinguishing feature of episodic memory. Meta-psychological autobiographical episodic memory from the inside retains a past perspective in an ego-like way, providing a conscious representation of one's psychological past. Insofar as such memory is consciousness of *self*, it involves consciousness of an individual's *conscious* psychological past.

The consciousness in the remembering is, I think, both phenomenal consciousness and rational-access consciousness.[34] That is, the episodic memories have a what-it-is-like quality to them; and, further, they are accessible to the individual's rational powers—propositional attitudes. Both sorts of consciousness are conditions of the whole individual, as distinguished from a psychological subsystem. Conscious psychological states that the relevant episodic memories are *of* are also both phenomenally conscious and rational-access conscious.

Perhaps not all individuals with psychologies are conscious. Some animals capable of perceptual representation may not be. But selves are conscious in both senses.[35] Phenomenal consciousness is the primitive center of conscious psychological life. Rational-access consciousness either incorporates phenomenal consciousness into propositional attitudes or consists in the individual's direct, occurrent control of propositional attitudes in mental agency.[36]

Imputation of activity or affect *to an individual*, as distinguished from attribution to the individual's psychological subsystems, motivates what has traditionally been taken as the aspect of an individual's psychology that is the individual's *own* in a proprietary sense. This notion applies, I think, to all individuals with a psychology. In higher animals, both sorts of consciousness are sufficient conditions for this proprietary ownership. They are constitutive of the complex type of psychology that determines what a self is.

---

[33] Kant assumes these points. In the subsequent discussion (*Critique of Pure Reason* A361–62), he takes the consciousness to be not only of self but of 'determinations' (*Bestimmungen*), which are properties of self that are of the sort that help make it a self—psychological properties.

[34] Here I use Ned Block's distinction, with a refinement that I have suggested See Block, "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences*, XVIII, 2 (1995): 227–47; and my "Two Kinds of Consciousness," in Block, Owen J. Flanagan, and Güven Güzeldere, eds., *The Nature of Consciousness. Philosophical Debates* (Cambridge: MIT, 1997), pp 427–34, also printed in *Foundations of Mind*, pp. 383–91; "Reflections on Two Kinds of Consciousness," in *Foundations of Mind*, pp. 392–419.

[35] No type of episodic memory is constitutively necessary for either phenomenal consciousness or rational-access consciousness. There are certainly phenomenally conscious beings that lack episodic memory. There are probably beings, higher animals and/or young human children, that have rational-access consciousness, but lack episodic memory.

[36] "Two Kinds of Consciousness"; "Reflections on Two Kinds of Consciousness."

I believe that the sort of consciousness that Kant had in mind *and* the sort of episodic consciousness that I have been developing support a more specific point about the consciousness of self over time. The consciousness is direct and noninferential—hence *de re*. Kant would not have counted conscious theory about one's past, based on *inference*, sufficient for *consciousness of self* in the relevant sense. Nor would he have counted memory that is a mere summation of past experiences *consciousness of self* in the relevant sense.[57]

Meta-psychological, autobiographical *episodic* memory *from the inside* meets this requirement of a *de re* relation to the past—and illuminates it. In meta-psychological uses, episodic memory represents *instances* or *tokens* of past psychological events or states. This sort of memory contrasts with two other kinds of meta-psychological autobiographical memory. One type is generalized memory of types or patterns that is an amalgam or summation of individual past experiences, but that does not retain any one past psychological event or state instance and represent it as it was at a specific past time. Memories with the following contents are examples: my experiences with La Tâche have been thrilling; I have often seen him dominate her; I have learned that she is not to be trusted. The other type is memory that characterizes a specific event, but in a generalized way that does not involve *de re* memory *of* the event. Memories like the following are examples: I have once seen a nut buried there; I have felt this sort of pain once before; at some point, I came to believe that Schubert's string quintet is sublime.

None of these memories is in itself a memory *of* a particular event instance. The memories either summarize several past events without necessarily remembering one of them as it occurred, or they retain *that* a particular event occurred without necessarily remembering that event as it was at a specific past time. Episodic memories are not only noninferential. They are directly *of* the remembered entity. They are *de re*. I think that meta-psychological, autobiographical, episodic memories from the individual's past perspective provide the sort of *self-consciousness over time* that, on the Kantian conception, is constitutive to being a self.

---

[57] Kant takes the consciousness to be via inner *sense*, which is noninferential: *Critique of Pure Reason* A361–63. I agree with Kant that the consciousness is noninferential. I do not agree that it must be via a *sense*—a capacity that yields only empirical cognition See note 1.

I turn now to why the diachronic aspect of this self-consciousness matters. What values attach to a self conceived in this way? I think that the conception figures centrally in the applicability and application of norms that are constitutive to the natures of selves and persons.

Locke took persons to be approximately what I call selves.[38] He characterized persons as thinking, reflective, self-conscious beings capable of extending consciousness to different times and places.[39] Locke regarded person as a 'forensic' notion, one essentially associated with responsibility.[40] Locke's first characterization of persons probably motivated Kant's focus on diachronic self-consciousness. Locke highlights extending consciousness 'backwards' in time to 'action or thought'. (See notes 39–42.) This emphasis is motivated by a concern that persons be able to understand their responsibility for past misdeeds for which they could be punished.

Locke gives no account of extension of consciousness into the past, although he clearly takes instances of such extension to be memories. His discussion of accountability, combined with his writing of *repeating* past consciousness in extending consciousness into the past,[41] suggests that he may have had episodic memory in mind. I shall assume that he did.[42]

---

[38] Locke took persons to be a subspecies of what *he* counted selves. He explained selves thus: "*Self* is that conscious thinking thing, (whatever substance, made up of whether spiritual, or material, simple or compounded, it matters not) which is sensible, or conscious of pleasure and pain, capable of happiness or misery, and so is concern'd for it *self*, as far as that consciousness extends." *An Essay Concerning Human Understanding* (Oxford: Clarendon, 1975), Bk. 2:27:17. See also 2:27:25–26. I call 'egos' *roughly* what Locke calls the selves of beings that are not persons. I do not require an ego to be conscious. Egos must be purposive and have some representational competencies, at least perceptual ones. Locke's persons are selves (in his sense of 'self') that are intelligent, have rational reflection, can extend their consciousness to different times and places, and are morally accountable. Locke's persons are *roughly* what I count selves. I say 'roughly', here and above, partly because I do not agree with Locke on the identity conditions for persons, selves, or egos. Some of these differences will emerge shortly.

[39] Locke's fuller characterization of person: "...a thinking intelligent Being, that has reason and reflection, and can consider itself as itself, the same thinking thing in different times and places, which it does only by that consciousness which is inseparable from thinking, and as it seems to me essential to it...in this alone consists *personal identity*, *i.e.* the sameness of a rational being. And as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that *person*; it is the same *self* now it was then; and 'tis by the same *self* with this present one that now reflects on it, that that action was done." *Ibid.*, 2:27:9.

[40] For the claim that 'person' is a forensic term, see *ibid.*, 2:27:26. Locke develops his views on persons and moral responsibility in 2:27:18–26. For an illuminating discussion, see Patricia Kitcher, "Two Normative Rules for Self-Consciousness," in Terrace and Metcalfe, eds., *op. cit.*, pp. 174–87.

[41] In *An Essay Concerning Human Understanding*, 2:27:10, Locke writes of persons' *repeating* "the idea of any past action with the same consciousness it had of it at first." This passage makes a relatively close fit with meta-psychological autobiographical episodic memory.

I believe that Locke was right to think that the relation between moral accountability and episodic memory is closely connected to the psychological *kind* that selves instantiate. However, he took the connection to hold at too specific a level. I will discuss these matters in more detail in the next lecture, but I sketch some issues here.

An individual can be morally responsible for deeds that he or she does not remember. An individual can be responsible for deeds that are motivated from a perspective that he or she does not know introspectively at the time the motivation is formed. Contrary to Locke, a drunkard can be morally responsible for misdeeds that are neither remembered nor understood at the time.[43] People mired in self-deception or naiveté about their own psychologies are similar cases in point. Moreover, longevity of life can lead to forgetting misdeeds for which one remains responsible. It is simply not true that one must have an autobiographical episodic memory, or indeed any other memory, for each past deed for which one remains morally responsible.

The connection of accountability with meta-psychological, autobiographical episodic memory holds at a more general level. As a first approximation, I think that having a natural competence consciously to revisit some actions and state instances from the inside is a competence that is constitutive to our *status* as morally accountable beings.

Darwin points in the same direction. But like Locke, he underestimates the subtlety of the connection between memory and being subject to moral norms. He writes:

> A moral being is one who is capable of reflecting on his past actions and their motives—of approving of some and disapproving of others,

---

[42] It has been disputed whether Locke understands 'extension' of consciousness back in time to be a type of memory. Some have maintained that he meant merely continuity of consciousness. I think that his taking this extension to be something a person does (*ibid*, 2:27:9–10) and his viewing personal identity and moral accountability as hinging on memory (*ibid.*, 2:27:20–22) suggest that he views extension of consciousness into the past as memory. At least, he takes memory to be a primary *type* of such extension.

[43] Locke's claims, now widely rejected, occur in *ibid.*, 2:27 22–26. Leibniz, *New Essays on Human Understanding*, Peter Remnant and Jonathan Francis Bennett, trans. and eds. (New York: Cambridge, 1981), 2:27:9, criticizes Locke's views on accountability and personal identity. Leibniz allows for memory gaps in both He maintains, correctly, that testimony of others can suffice to enable one to understand the justice of punishment for unremembered deeds. Of course, it does not follow that if all that one ever had had—even for very short-term time spans—was testimony from others about one's past, one would be a morally responsible person.

> and the fact that man is the one being who certainly deserves this designation, is the greatest of all distinctions between him and the lower animals.[44]

Although Darwin shares Locke's focus on memory, he aims—fruitfully, I think—at a generic connection between memory and moral status, rather than a connection between memory and each morally responsible act. On the other hand, Darwin is less focused on episodic memory than Locke is. Furthermore, approval and disapproval are not enough to make an individual a moral being, even when added to diachronic competencies that are relevant to moral status. An individual must also understand a distinction between right and wrong. Nevertheless, I think Darwin right to connect diachronic competencies with being subject to moral norms.

The constitutive connection between being subject to moral norms and being competent to extend consciousness over time is not that meta-psychological autobiographical episodic memory is the link that enables a person or self to remain the same person or self over time, as Locke seems to have thought. Identity is a necessary condition on primary moral accountability. But episodic memory is not the primary link in trans-temporal identity. There are many other psychological links, including unconscious ones and ones other than memory, that help make possible the identity of selves and persons over time.

The connection between diachronic competence and moral accountability is more abstract. In lacking a competence for trans-temporal 'extensions of self-consciousness'—of which a competence for meta-psychological autobiographical episodic memory is a central example—an individual lacks an essential aspect of a certain *understanding*. A capacity to understand one's actions in this way is necessary for being subject to moral norms, and partly constitutive of being a self. My next task is to explore the kind of understanding that has these features—with special attention to the role that memory plays in this self-understanding.

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA
ALIPORE CAMPUS

---

[44] Charles Darwin, *Descent of Man* (1871) (Amherst, NY: Prometheus, 1998), p. 633.

SELF AND SELF-UNDERSTANDING

LECTURE II: SELF AND CONSTITUTIVE NORMS

In this lecture, I pursue a certain methodological strategy. I want to contribute to understanding the nature of selves. I think that certain sorts of self-understanding are among the constitutive features of selves. I explore that type of self-understanding by using clues from requirements on the applicability of two types of norms—those of critical reason and morality. Being subject to norms of critical reason is plausibly constitutive to being a self. Being subject to moral norms is at least arguably constitutive. Having certain psychological competencies, including relevant sorts of self-understanding, is constitutive to being subject to these norms. So the requirements for being subject to these norms provide insight into the nature of the relevant self-understanding, and ultimately into the nature of selves and persons.

I

A *norm* is a standard that governs good or optimal ways of fulfilling some function, or achieving some goal, end, or value.[46] Many norms do not require for their applicability any sort of understanding on the part of individuals that are subject to them. Norms for having a given standard of living or for having a life-sustaining nutrition, for any organism, are applicable even if they are not understood. Norms for the well-functioning of the heart or for formation of veridical visual perceptions require no understanding. They apply whether or not anyone knows about them. Norms are applicable given any function. Norm is a teleological notion. Not all teleology involves psychology, much less understanding.

Even some *rational* norms governing propositional attitudes require neither self-understanding nor understanding the norms. Reasoning by young children is subject to norms of deductive and inductive inference. Being subject to these norms requires no understanding beyond competence to make such inferences. An individual is subject to these norms if it engages in inference. Inference has the representational function of preserving truth, supporting truth, or supporting some goal. Norms of reasoning apply to any being with propositional attitudes. Organisms that lack propositional attitudes can be evaluated for efficiency, but not reasonability. But young children and probably some nonhuman animals are subject to norms of reason, even though they cannot understand the norms.

---

[46] See my "Perceptual Entitlement," *Philosophy and Phenomenological Research*, LXVII, 3 (November 2003): 503–48; and "Primitive Agency and Natural Norms," *Philosophy and Phenomenological Research*, LXXIV, 2 (September 2009): 251–78.

I shall center on moral norms and norms of critical reason. Moral norms are norms for acting morally well, for making good decisions, for reasoning well to moral conclusions, and for having good character. Norms of critical reason perhaps need more explication.

Critical reason is a capacity to recognize and effectively employ reasonable criticism of or support for propositional attitudes and for propositional reasoning, guided by an appreciation, use, and assessment of reasons and reasoning as such. Critical reasoning is the sort of reasoning that evaluates, checks, refines ordinary first-level reasoning, using concepts and standards of reason. Critical reason is constitutively a capacity for meta-reasoning and meta-evaluation. It is commonly reasoning about ordinary subject matters and about reasons at the same time. It evaluates attitudes and reasoning for their reasonability.

Most of our reasoning, even most of our most sophisticated reasoning, proceeds at a first-order level. Most of our reasoning does not refer to representational contents or psychological states, and does not invoke the concept reason or standards for being reasonable. It simply operates under those standards more or less well. Critical reasoning is not in itself any better or more effective as reasoning than first-order reasoning. *Actual use* of critical reasoning is not necessary for making scientific discoveries. But a *capacity* to frame and defend a methodology is constitutive of science. Entering fully into practices that allow us to have our culture and history requires some capacity to explain and defend the practices. One must be able to answer why questions if one is to stand behind—with good or spurious reasons—the values, taboos, and permissions, the institutions and relationships, that make up a civilization. Such ability constitutes a rudimentary capacity for critical reasoning.

We are probably not the only species that reasons. Other animals probably carry out propositional, means-end inferences, subject to norms of reason. Ability to reason is probably not distinctive of selves or persons. But I think that ability to reason *critically* is.

I will not rely on a view about the relation between norms of critical reason and those of morality. Some, like Hume, hold that moral values are a matter of nonrational choice—a rationally optional project.[46] Others, like Kant, hold that moral values are matters of reason. Some think that a person or self could be irretrievably and naturally morally blind or indifferent—without damage, deprivation, illness, or malfunction. Others, following Aristotle and Kant, maintain that moral values are associated with constitutive functions of human beings, persons, or selves. I am more sympathetic to the latter in each pair of views. Since the self-understanding that I discuss is parallel for morality and critical reason, little in the discussion will depend on either of the issues cited in this paragraph.[47]

---

[46] Even Humeans can allow a constitutive role for critical reason in morality—a role for a capacity to defend choices, including moral ones, as reasonable, *relative* to one's moral values.

I think it hardly controversial that having a natural capacity for critical reason is constitutive to being a person and being, or having, a self. It is arguable that being subject to moral norms is also constitutive. I will focus on both kinds of norms as clues to the nature of a type of self-understanding that is constitutive to being a person and self.

An individual's instantiating certain psychological kinds—including a kind of self-understanding and an understanding of notions like *reason* and *moral wrong*—is constitutively necessary, and arguably sufficient, for the applicability of the norms of critical reason and morality to that individual. Reciprocally, these psychological kinds necessitate the applicability of certain norms to their instantiations. So psychological capacities and normative applicability are interdependent. I shall use the applicability of these norms as clues to the natures of underlying psychological capacities.

The norms of morality and critical reason are not typical norms. Unlike norms for good nutrition, cardiovascular health, and veridical visual-perception formation, these norms apply only to beings with certain psychological capacities for propositional attitudes. The norms' applicability hinges on two other requirements. They require *understanding the norms*—some minimum internalization of them. And

---

[47] Locke's claim that *person* is a forensic notion expresses his view that there is a constitutive connection between being a person and being responsible to moral norms. This view is independent of Kant's more committal view that morality rests on critical reason. Hume differs from Kant in holding that morality does not depend on reason's determining basic values. I do not know whether he thought that a capability for morality is necessary to being a nondamaged fully mature person, or human being.

It is not uncommon to take Locke's claim to imply that the kind *person* does not mark off individuals and is not a basic psychological kind. Locke understood a forensic notion as one having to do with judicial proceedings and law, including divine moral judgments. It is, even now, common to contrast forensic notions with "ontological" notions like *substance*. Such a contrast is often taken to imply that the kind *person* marks a stance, phase, or contingent property—not a kind of individual. I believe that such contrasts often depend on an outmoded notion of substance. The stance-oriented way of thinking about normative notions also tends to conceive norms as applicable only given desires or choices There are some norms of this sort. Perhaps norms of table setting apply only if one cares to conform. But norms of critical reason are not like that. And neither, I think, are norms of morality. I believe that these norms apply to individuals if and only if the individual has certain psychological capacities and certain rudimentary understanding At any rate, I think that counting *person* and *self* notions with normative implications does not *entail* that they apply to less ontologically significant kinds they require some *self-understanding*—an understanding of psychological conditions, in one's own case, to which the norms apply.[48]

Let us briefly consider the requirement of understanding these norms.[49] Being subject to moral norms requires some understanding of the difference between right and wrong. Both right and wrong are moral, normative notions. Right and wrong centrally concern psychological motivations and psychological initiators of action. The motivations and initiators must involve conceptualization of actions. Application of the norms to actions requires that actions be individuated partly by their psychological antecedents. Helping an old lady across the road mainly motivated by a desire that she arrive at the bank where a scam is planned is a morally different act from helping the old lady across the road mainly out concern for her well-being.[50] Minimal understanding of how moral notions apply requires not only *having* such notions, but being able to use them to evaluate psychological conditions.

---

[48] Tyler Burge, "Reason and the First Person," in Barry C. Smith, Crispin Wright, and Cynthia Macdonald, eds., *Knowing Our Own Minds: Essays on Self-Knowledge* (New York: Oxford, 1998), pp. 243–70.

[49] Development in children of a capacity to apply norms is not deeply understood. Here is a rough summary. Child development from ages three to six is marked by changes in understanding the role of norms in self-regulation. By age three or four, when children clearly exhibit an ability to attribute beliefs and intentions, they show some awareness that others have different beliefs and goals. Between ages three and six, children exhibit an understanding that social regulation by others is mediated by others' attitudes. Evaluations had been applied to appearance, physical prowess, social behavior. They come to be applied to psychological states. Evaluation is initially accepted from others. It comes to be self-applied. Between ages six and 10, children come to evaluate individuals' psychological conditions and characters not just in terms of whether an outcome matches an antecedent intention or belief. They evaluate individuals in terms of the quality of the antecedent state, somewhat independently of whether an outcome is favorable. This distinction probably grows out of a distinction between talent and effort in physical pursuits. Such evaluations have a three-factor profile. First, there is the contribution of the individual's abilities (those relevant to talent, character, psychological capacity) to whether a standard is met. Second, there is the contribution of the individual's use of these abilities. Third, there is the contribution of external conditions to success or failure. Moral evaluations of individuals and their acts center on psychological capacities, character, and use of them. An individual can act well morally, although things turn out badly. An individual can act badly, although things turn out well. Similarly, evaluations for critical reason center on the individual's contributions, not on extra-psychological matters. Such evaluations are usually independent of the veridicality of an attitude, and of the success of the activity in realizing the individual's goals. An individual can be rational but mistaken, or right but irrational.

For a sampling of literature on development of cognition of norms, see Henrike Moll and Tomasello, "Level 1 Perspective-Taking at 24 Months of Age," *British Journal of Developmental Psychology*, xxiv, 3 (September 2006). 603–13; Charles W. Kalish and Sean M. Sheverick, "Children's Reasoning about Norms and Traits as Motives for Behavior," *Cognitive Development*, xix, 3 (July–September 2004): 401–16; Nancy Eisenberg, Richard A. Fabes, and Tracy L. Spinrad, "Prosocial Development," in Eisenberg, William Damon, and Richard M. Lerner, eds., *Handbook of Child Psychology, Volume 3: Social, Emotional, and Personality Development* (New York: Wiley, 1998); Adam Rutland et al., "Social Norms and Self-Presentation: Children's Implicit and Explicit Intergroup Attitudes," *Child Development*, lxxvi, 2 (March/April 2005): 451–66; Gergely Csibra and György Gergely, "Social Learning and Social Cognition: The Case for Pedagogy," in Yuko Munakata and Mark H. Johnson, eds., *Processes of Change in Brain and Cognitive Development: Attention and Performance XXI* (New York: Oxford, 2006), pp. 249–74; David R. Shaffer and Katherine Kipp, *Developmental Psychology: Childhood and Adolescence* (Belmont, CA: Wadsworth, 2009)

Similarly, being subject to norms of critical reasoning requires minimal understanding of norms of critical reasoning—understanding what a reason is. The minimal understanding involved in having the concept reason requires understanding applications of the concept to *instances* of reasoning.[51] Reasons apply centrally to representational contents taken with their mode (belief, suspension of belief, intention). But, as noted, understanding the concept reason requires a capacity to apply it to instances. Understanding conditions under which the concept applies in actual reasoning requires a capacity to apply it to psychological states—instances as well as types—as individuated by their modes and representational contents.

Thus to be subject to either moral norms or norms of critical reason, an individual must have relevant normative concepts and abilities to apply such concepts competently to psychological states and occurrences, and their representational contents, as such. These are *higher-order*, ability-general meta-psychological capacities.[52]

My main interest here is the requirement, for the applicability to an individual of norms of morality and critical reason, that the individual have self-understanding. To be subject to norms of morality, one must understand application of concepts of right and wrong to psychological states or to actions individuated in terms of psychological antecedents. This understanding must be applicable to one's own case. One must be able to evaluate one's own psychological states, particularly inasmuch as they determine and help individuate one's bodily actions; and one must be able to act on such evaluation. These abilities require an ability to understand what one's psychological states are.

Similarly, norms of critical reason are applicable to an individual only if the individual has a standing competence, at least in some

[50] I believe that I disagree here with Derek Parfit and Tim Scanlon, both of whom minimize motive in determining the moral goodness of an action. See Derek Parfit, *On What Matters*, vol. 1 (New York: Oxford, 2011); and Thomas M. Scanlon, *Moral Dimensions Permissibility, Meaning, Blame* (Cambridge: Harvard, 2008). For a view that I find more congenial, see Barbara Herman, "A Mismatch of Methods," in Parfit, *On What Matters*, vol. 2, ed. Samuel Scheffler (New York: Oxford, 2011), pp. 83–115.

[51] Indeed, instances of reasoning in one's own case. See my "Reason and the First Person."

[52] For explanation of the ability-general/ability-particular distinction, see my "Five Theses on *De Re* States and Attitudes."

cases, to understand his or her point of view and to criticize and affirm that point of view by reference to reasons considered *as* reasons. In exploring these requirements, I shall be guided by the idea that there must be a capacity to understand one's point of view and the applicability of normative concepts to it "*from the inside*." Like memory from the inside, understanding from the inside must get the understood state's representational features right, and must do so by preserving those features in certain ways. Preservation is a fundamental connector in any point of view. Selves capitalize on this connector. Conditions for applicability of norms of critical reason and morality center on understanding that makes essential use of preservation.

These points about baseline conditions for the *applicability* of the norms are fairly straightforward, as far as they go. I think that we can gain insight into the specific nature of the required self-understanding by reflecting on conditions for the *application* of the norms in particular instances. The patterns of evaluations in such applications can be very complex, especially in the moral case. But some simple points about how antecedent psychological states and occurrences figure in such applications will, I think, lead us a step deeper.

II

Let us first look at some applications of moral norms. Moral norms apply most conspicuously to an individual's acts individuated in terms of psychological antecedents. Individual responsibility for such acts is centered on psychological antecedents. An individual can be just as responsible for a bodily act as for psychological antecedents that initiate it. But in such cases full, unmitigated responsibility for the act derives from the act's proceeding from its psychological antecedents in a natural way. The root of responsibility for the act is the individual's responsibility for psychological antecedents—the cognitions, intentions, decisions, and psychological initiations—that are subject to a certain sort of self-understanding. Moral responsibility has its constitutive, explanatory basis in psychological antecedents of acts, antecedents that are subject to a certain sort of self-understanding.

Similarly, an individual is morally responsible for consequences of his or her acts only if the consequences are traceable to something that the individual intentionally did or omitted, where the individual was capable of understanding the intention and initiation of his act, or omission, from the inside, and recognizing the act's moral implications.

On the other hand, an act can be wrong and an individual can be morally responsible for it, even if the individual does not specifically intend, or could not foresee, its harmful consequences. An individual can be morally responsible for acts carried out at times when the individual cannot understand what he or she is doing, in morally relevant terms. I noted the classical case of the drunk last lecture. An individual wild with anger can be morally responsible for injuring another, even if the injury was not specifically intended and even if the perpetrator was not in a state to understand the recklessness of the behavior. The individual might not even realize that he or she was angry or had reckless intentions. The individual could still be morally responsible for the behavior and the injury—not just for not knowing what he was doing and not just for not foreseeing the injury. Similarly, arrogant behavior is often not backed by specific intention or self-understanding. Racist beliefs are often not understood to be racist. Insensitive ingratitude or sexist acts are often not intended or understood as such. Yet an individual can be morally responsible for these states or acts.[63]

Criticism of the acts and the individual is often *mitigated* in such cases. The drunk driver is not morally as bad as the driver who lies in wait. An inappropriately angry person who would be horrified by his act in a cooler moment is often less bad than one who calculates injuring. Individuals are not strictly morally liable for all causal consequences of their acts. But neither are they always absolved of moral responsibility because they lack the self-understanding to evaluate them as they are formed.

I think that self-understood psychological antecedents in carrying out acts do, however, have a special place in *application*—not just *applicability*—of moral norms. The pattern of mitigation just cited suggests this point.

Responsibility by the drunk person hinges on the relation of the act to a baseline condition in which the individual controls his or her intentional acts and is in a position to understand their psychological antecedents and the acts' moral significance. At some time before the driver became drunk, perhaps in the development of drunken habits, the individual must have had a capacity to foresee risks to others, if the individual is to be counted morally responsible for the act. At some such time, the drunkard was in a position to understand and evaluate psychological antecedents to acts that in some foreseeable way led to being drunk.

Similar points apply to the cases of anger, arrogance, insensitive ingratitude, racism, and sexism. If these characteristics were to have developed in such a way that at *no* point was there *any* capacity to understand, evaluate, and control relevant psychological antecedents, the individual would not be morally responsible for them. It would be as if the states were injected into the individual. We control only little of our psychologies at any given time. But over time, we have opportunities to shape them. Such opportunities make possible moral responsibility for more than we control or understand at any given time.

---

[63] Robert Merrihew Adams, "Involuntary Sins," *The Philosophical Review*, XCIV, 1 (January 1985): 3–31. Several of my examples are drawn from Adams's provocative article. He argues against always tracing responsibility in these cases to earlier culpable willings and understandings. I am not persuaded

The mitigation in moral evaluation of the drunk driver in comparison with the calculating murderer derives from a distance between the driver's condition at the time of the accident and the driver's condition when he was in a position to understand what he was doing. The fact that the driver had less understanding of and control over his point of view and its moral consequences at the time of the accident than at an earlier time is *commonly*—I do not say 'always'—a source of mitigation. Moral evaluation of individuals ultimately centers on ways an individual's acts and states relate to uses of the individual's point of view that *are* under intentional control and capable of self-understanding. Moral evaluation takes such uses as a baseline for determining mitigation.

The *application*, as distinguished from *applicability*, of norms of *critical reason*—especially theoretical critical reason—is less complex than application in the moral case. But again, ultimately, certain elements in an individual's point of view are privileged in evaluations. Let me lay out some of these differences and similarities step by step.

Norms of critical reason include norms of first-order rationality—those that apply to any reasoner—together with those rational norms that are specific to critical reasoners. Norms specific to critical reason apply only to propositional states and events that are in principle accessible to an individual's rational powers, immediate self-understanding, and rational evaluation, using the concept reason or some variant. Other aspects of an individual's psychology can be unreasonable, but not specifically *critically* unreasonable. However, elements of a point of view that the individual does not control, access, or understand fall under norms of critical reason. Elements in the Freudian unconscious, elements mired in self-deception or inebriation, elements distorted by entrenched irrationality, and elements camouflaged from self-understanding by anger or depression are subject to norms of critical reason. In this respect, application of moral norms and norms of critical reason are similar.

A difference is that norms of critical reason apply with equal force to all elements of an individual's point of view to which the norms are applicable. In the moral case, intentions and actions not presently understood or available to self-understanding are often *not* just as morally bad as they would be if they had been understood. The norms of critical reason include norms of ordinary non-meta-representational reason. Unreasonable elements in a point of view that are not presently available to self-understanding and self-evaluation are just as unreasonable as they would be if they were. The drunkard's inferences, which are not immediately accessible, but fall under standards of reason, are just as unreasonable as a sober person's making the same inferences under reflectively ideal circumstances. Some psychological states or transitions that would normally be open to self-understanding might be influenced by disease. Their unreasonability need not be explained in terms of something the *individual* did or did not do in allowing them to lapse, or in failing to bring them under the control of critical reason. Unlike immorality, unreason—even specifically critical unreason—can be *purely* the product of disease or malfunction. In assessing *failures* of reason, there seems to be no privileging of elements that are presently open to self-understanding over those that could be open to self-understanding, but are not.[54]

Still, as in the moral case, reasoning that is easily accessible to current self-understanding is privileged in *successes* of critical reason. Successes are constitutively open, readily open, to critical reflection that includes self-understanding. Psychological elements that are not currently easily accessible to self-understanding can fulfill norms of critical reason only via inertia: they were readily open, and have maintained their reasonability.

An individual can use critical reasoning to operate on elements of the point of view that are resistant to critical reason, and—although they may be in principle accessible—may not be currently, easily accessible to self-understanding. We say that the individual *tries to get him- or herself* to think or act more reasonably. The recalcitrant elements are not, however, part of the critical reasoning. They are within the individual's point of view, and are subject matters of higher-order reasoning. But the aspects of the point of view that are the topics of the reasoning are not being used in critical reasoning. They are like objects to be manipulated. The resources of successful critical reasoning constitutively involve elements that are currently, easily accessible to self-understanding.

[54] There is, however, a pattern of mitigation that resembles the moral case. Suppose that an individual's reasoning is periodically clouded by anger Suppose that the individual tries to change matters for the better, but does not succeed. The unreasonable states are just as unreasonable as if the individual had done nothing to combat them. But the *individual's* critical unreasonability can be mitigated The *individual* is perhaps less critically unreasonable than if he or she gave in to the unreason.

Meeting the norms of critical reason and morality occurs paradigmatically—I think, only—in cases identical with or derivative from cases in which an individual is currently in a position to readily understand the relevant norms and the application of the norms to relevant psychological states and occurrences. Understanding such application requires that the psychological elements be currently, readily accessible to self-understanding.

Let me summarize my points about applicability and application of the two types of norms. Both types center on psychological elements that are imputable to the individual subject to the norms—to that proprietary part of the individual's psychology in the individual's point of view. What happens in the environment, in intentional bodily actions, or in aspects of the individual's psychology that are irretrievably beyond the individual's control or consciousness, or that do not accord with the individual's self-understanding, have a fundamentally different status in the norms' applicability and application than elements within the individual's point of view that are subject to self-understanding.

III

Animals that are not persons and lack selves have points of view. Even in animals' points of view, there is a distinction between elements that are occurrently conscious, or are immediately or easily accessible to consciousness—whether phenomenal consciousness or rational-access consciousness—and elements that are not. Thoughts or perceptions that are retained in memory, but that would need extensive priming to gain consciousness or be available for rational use, fall in the latter category. Standing beliefs or memories that are available for use and control as soon as the occasion requires fall in the former category.

I call the part of an individual's point of view that is rational-access conscious, or immediately or easily accessible to such consciousness, '*the individual's rationally accessible point of view*'. Elements not immediately or easily accessible but imputable to the individual are part of the individual's point of view, but not part of the individual's rationally accessible point of view. Perceivers that lack propositional attitudes have a point of view, but no rationally accessible point of view.[55] Probably all individuals with rationally accessible points of view have further elements in their points of view, outside the rationally accessible point of view. They probably have memories that need priming to come into play.

---

[55] In individuals with a rationally accessible point of view, there is a further distinction among elements of the point of view that are outside that part of their point of view. There are states imputable to an individual outside that are not accessible to consciousness even with priming, as well as those that are. Imputability to the individual is not fundamentally a matter of accessibility-in-principle to either sort of consciousness. Individuals have short-term, unconscious perceptual beliefs and memories that enter

For persons and selves, this distinction between point of view and rationally accessible point of view has a higher-level analog. There is a distinction between an individual's representational states and events that are immediately or easily accessible to conscious *self-understanding*, on one hand, and representational states and events that are imputable to the individual, part of the individual's point of view—and perhaps even in principle accessible to self-understanding—but not immediately or easily accessible, on the other. I call the former representational states and occurrences, together with conscious sensations and feelings that are also immediately or easily accessible to self-attribution, the individual's '*apperceptive rationally accessible point of view*'.[56] An individual's unrecognized anger, Freudian unconscious, self-deception, and other states that need extensive priming to be recognized, are outside the individual's apperceptive rationally accessible point of view.[57] The apperceptive rationally accessible point of view is the part of a point of view that is immediately or easily accessible to conscious self-understanding.

Our reflection on applicability and application of moral norms and norms of critical reasoning indicated that apperceptive rationally accessible points of view have a privileged status. Applicability and fulfillment of these norms require a capacity for self-understanding

---

into individual-level explanations of an individual's activity, but that are forever beyond conscious control or phenomenal awareness. I think that such beliefs are imputable to the individual, first, because all beliefs are imputable and, second, because they can enter into individual-level explanations of action. This point will have an analog in discussion of critical points of view: some psychological states that are imputable to an individual cannot be claimed by the individual, even in principle, as the individual's own and are not under control of the individual's agency. (Of course, no such states are *morally* imputable.) In other words, psychological imputability and point of view have their roots in committal states and activity not in consciousness. Committal perceptual states and activity are present in the animal world at very low phylogenetic levels. These points are partly terminological. But terminological choice has important substantive implications for how we think about psychology. Whatever the terminology, these distinctions should be drawn.

[56] In persons and selves, anything in the apperceptive rationally accessible point of view is in the rationally accessible point of view. The converse may also be true: anything in the rationally accessible point of view is, in persons and selves, in the apperceptive rationally accessible point of view. I leave this issue open to further investigation.

[57] *Occurrently* conscious elements in apperceptive rationally accessible points of view have a priority in our *sense* of life. I leave open whether they have normative priority. For purposes of the large normative considerations that we are discussing, it is the full rationally accessible point of view that is privileged.

of relevant psychological elements and a capacity to evaluate such elements under the norms. Apperceptive rationally accessible points of view set standards for successful realization of the norms. They form a baseline for mitigation in negative moral evaluation.

I term this fundamental status of apperceptive rationally accessible points of view in explanations of normative evaluation in morality and critical reason their '*buck-stopping status*'. There are no further or deeper grounds for evaluating individuals' contribution to their accountability under norms of morality or critical reasoning. Applications of these norms are rooted in individuals' apperceptive rationally accessible points of view.[58]

The idea behind this point is that elements in a point of view that have buck-stopping status with respect to norms of morality and critical reason are constitutive starting points for expression of an individual's moral being or critically reasonable self. They mark ultimate unmitigated responsibility to the norms. All other elements in an individual's point of view are at best mediate expressions. They get their normative status derivatively—through relations to the apperceptive rationally accessible part of the point of view.

The primary relations here are inferential ones. The buck-stopping status of apperceptive rationally accessible points of view lies largely in their being baseline starting points for inferences. These baseline starting points necessarily are immediately, noninductively accessible, at least within the conscious apperceptive point of view. 'Easiness' of access is meant to be unspecific. Easiness is relative to the starting points that are immediately accessible to self-understanding. In having easy access to some elements in a point of view by inference, one must self-understand the starting points of the inference.

Although both moral evaluation and evaluation of critical *practical* reason usually center on bodily intentional action, the baseline, buck-stopping starting points for such evaluation are the psychological antecedents of the actions. Let me make this point somewhat more precise. I shall henceforth call intentional bodily action just '*action*'. As noted, one can be equally accountable under the two types of norms for action and for its psychological antecedents. The capacity to reflect on one's actions, especially prospectively, is a factor in relevant normative evaluation. My claim is that the applicability and application of the norms of morality and critical reason to actions goes through their applicability and application to the psychological antecedents of the actions. The individual's contribution to his or her accountability for actions derives from the actions' relations to self-understandable cognitive and motivating psychological antecedents. Full accountability for actions depends on ability to self-understand psychological antecedents.[59] Where understanding actions themselves is a source of the individual's contribution to his or her accountability, the understanding of the actions must be through understanding their cognitive and motivating psychological antecedents. Basic, full, unmitigated accountability for an action depends on the action's realizing the content of those antecedents, where the antecedents can be self-understood. If the action is not individuated by the self-understandable psychological antecedent—for example, if it does not realize an antecedent intention—the evaluation of the act is mitigated, or at least depends partly on its relation to actions that *are* successful expressions of self-understandable psychological antecedents. The starting point and ultimate source for evaluation of individuals under the two norms is the set of cognitive or motivating psychological antecedents. These help individuate the actions that are evaluated. They obtain their buck-stopping status partly through their potential for being self-understood.

---

[58] I think that not all self-understanding that grounds accountability under the relevant norms is agential. I discuss relations between rational-access consciousness—the lower-level underpinning of an apperceptive rationally accessible point of view—and individual agency in "Two Kinds of Consciousness."

The converse points do not hold. The basis of accountability does not lie in understanding the cognitions and motives through understanding the individual's actions. We often do understand our motives retrospectively, from the actions inward. But such understanding does not ground *fundamental* accountability of individuals under the relevant norms. Motives that are understandable only outside-in are less fundamental expressions of an individual's moral and critically rational being. And actions that derive from such motives are often evaluated as involving mitigated accountability. Basic, unmitigated accountability traces to cognitions and motives that are subject to potential self-understanding in advance of, or during, actions that they produce. The *fundamental* basis for evaluating individuals' contribution to accountability under the two primary norms is the self-understandable cognitive and motivating psychological antecedents, not the self-understandable actions.[60]

---

[59] There is some question about whether individuals always know directly the initiation of a bodily action—when and whether the final choice or action initiation is made. Cf. Masao Matsuhashi and Mark Hallett, "The Timing of the Conscious Intention to Move," *European Journal of Neuroscience*, xxviii, 11 (November 2008): 2344–51. As long as the individual is capable of understanding and knowing, in a relevantly direct way, the motivating intention, the conditions on self-understanding of an action in doing it are met. If the action initiation flows naturally from an intention and decision and is not a surprise, the individual can be fully accountable for his or her action, under the relevant norms, if the intention and decision are open to self-understanding, even if the final occurrent action initiation is not The latter might well be unconscious

The ground for buck-stopping status of the *psychological antecedents* of action is that an action's nature or consequences may not conform to the individual's motivations, and may not be foreseeable by the individual. The individual's accountability depends on what he or she intends and is in a position to foresee. The specific sorts of accountability that are the focus of the norms of morality and critical reason depend on those aspects of an individual's point of view that are, at some point, accessible to self-understanding.

Self and hence *person* are kinds that are constitutively determined by capacities for self-understanding and for self-applying certain norms. Aspects of a self or person marked by these starting points are *the core self*. Core selves are partly individuated by their rationally accessible points of view. I want to understand the nature and normative implications of the *self-understanding*—distinguished from the self-application of norms—that is constitutive of core selves.

IV

Let us return to the Kant/Locke conception of a person or self as constitutively having *diachronic* powers of self-consciousness. Kant gives little explicit motivation for his conception (see note 1). Locke provides a motivation that claims too much.

In the previous lecture, I discussed meta-psychological autobiographical episodic memory from the inside as a prime instance of a diachronic self-consciousness constitutive to being a self. I want to explain how this type of memory and an analogous type of anticipation figure in competencies constitutive to selves. I discuss three ways in which the self-consciousness constitutive to selves—and to being subject to moral norms and norms of critical reason—has such diachronic implications: in inference, in practical decision making, and in dialectical aspects of critical reasoning and moral thinking.

---

[60] These points must be understood in the complex context of moral evaluation discussed earlier. Individuals are morally accountable for actions whose outcomes and even natures they did not intend, if those outcomes and natures were foreseeable. Some philosophers have broadened these points by appealing to a notion of moral luck. See Thomas Nagel, "Moral Luck" (1976), in *Mortal Questions* (New York: Cambridge, 1979). I think that more exact analysis shows that the phenomena that Nagel discusses are not as destructive of our notions of agency and moral evaluation as Nagel advertises. To avoid relying on any particular view of moral luck, I center on full, unmitigated accountability, and on the applicability and application of moral norms and norms of critical reason to the *individual's contribution* to accountability. I believe that in a sense the individual's contribution is what these norms are *all* about. But the notion of the *individual's contribution* must be understood, and applied, in a nuanced way to make out this "sense." In any case, I stand by the view that a competence for self-understanding of antecedent cognitive and motivating psychological elements grounds full, unmitigated individual accountability, and grounds normative evaluation of the *individual's contribution* to his or her accountability for actions, under the relevant norms.

The first way lies in inference. I make two assumptions. First, I assume that selves must be capable of propositional inference. Second, I assume that they must be able to think consciously *de re* of some of their propositional inferential events, as their own. The apperceptive rationally accessible point of view must reflexively apply *de re* to inferential events in it. Selves not only must be able to be conscious of *steps*, the premises and inferred contents, in inferences and be able to understand steps as reasons for conclusions. They must also be able to be conscious *de re* of events that make up the inference.

The first assumption is straightforward. As critical reasoners, selves must be able to make rational propositional inferences. The second assumption needs elaboration. Its requirement of consciousness is not what needs it. Being able consciously to carry out a self's constitutive representational functions helps define selfhood. The requirement that one think of inferences as one's own is also straightforward. What needs elaboration is the requirement that selves be able to think *de re* of their inferential events. Two considerations underlie this requirement.

One concerns the role of inference in the buck-stopping status of some of a critical reasoner's states. For any states in a critical reasoner to have such status, some inferences must have it. For being rationally accountable for state instances in one's point of view requires ability to support them by rational inference. Inferences, as token events, must be among the buck-stopping elements in a critically rational point of view. As a consequence of a point I made earlier about induction, if an individual could only infer inductively to what his/her inferences are, inferences could never be the starting point for the accountability of a critical reasoner. So one's access to one's inferences must not always be by inductive inference. Inductive inference is the only *inferential* access one *could* have to one's inference events. Moreover, the access must be through canonical—not purely descriptive—representations of the inferences: the access must be to the inferences on their own terms. Purely descriptive access would not allow the inferences, and their premises, to be starting points of accountability. A noninferential, not-purely-descriptive access is *de re*.

The other consideration behind the requirement of *de re* access to one's inferences concerns the role of such access in having a minimal competence with the concept of reason—a requirement on being a critical reasoner. Minimal competence requires understanding implementation of reasons as such—understanding being moved by reasons in particular cases—in the practice of one's own thinking.[61] If one could not implement reasons as such with respect to state instances that one has in mind—in the practice of one's own thinking—one could not understand implementation of reasons as such. If one lacked *de re* access to one's inferences—and knew them, if at all, only purely descriptively or by inductive inference—they would not be *part* of one's own critical thinking: one could not have them in mind. They would at most be topics of appraisal: like unconscious states that one knows one has, but cannot sustain or change in current critical reasoning. Implementation of reasons as such must be *de re*, and it requires *de re* access. If an individual knew only in a non-*de re* manner that he or she had engaged in an inference, the individual would be able to know whether and why he or she came to a conclusion and perhaps could appraise the inference in some way.[62] But the knowledge would not allow implementation of reason appraisals, any more than one could implement reason appraisals in others' thinking. The inference that the individual would know about, the topic of the knowledge, would not be an integral part *in* any critical reasoning. The inference could not be held in mind. So in the absence of *de re* access to his or her own inferences, an individual could neither implement reasons as such with respect to inferences in the practice of his or her own critical thinking, nor minimally understand implementation of reasons with respect to inferences—being moved by them as such. Selves are critical reasoners. Critical reasoners must understand implementation of reasons as such in their own inferences. So to be a critical reasoner, hence to be a self, one must have *de re* access to one's own inferences.

The two assumptions that I have just motivated, in two ways, can be used to show that selves must have meta-psychological autobiographical episodic memory from the inside. The first assumption of a capacity for inference entails a capacity for memory. Inference is not a point event. It requires some representational retention. Retention of mode and content through an inference is an instance of purely preservative memory.[63] Purely preservative memory is not autobiographical episodic memory from the inside. But the requirements on being a self in the second assumption entail that having more than purely preservative memory is required for the relevant inferential capacities.

---

[61] I have argued this point at some length in "Reason and the First Person," p. 250ff. I think that at least the first, and possibly the second, of the two considerations that I am expounding can be modified to show that having *de re* access to one's inferences is necessary to engaging in the minimal reasoning necessary to being morally responsible

[62] I think that one cannot engage in reason appraisal if one cannot implement reasons as such—be moved by reasons as such to change or sustain psychological elements in one's own thinking Cf "Reason and the First Person." Both appraisal and implementation are necessary to have the concept *reason*. It should be noted that the argument of this paragraph could dispense with this premise I have used it because it fits with the earlier discussion. The minimal point is that critical reasoning requires that one reflexively hold a state instance in mind as one deliberates with it, evaluates it, and determines according to reason evaluation whether to sustain it or change it. Doing so requires *de re* access to the state instance.

The second assumption, with its requirement of a capacity to think consciously *de re* of inferential events as one's own, immediately entails that the relevant inferential capacities carry a competence with *meta-psychological autobiographical* memory. *Episodic* memory must be long term, conscious, and of occurrences, as they were at a given past time. It is a form of experiential memory.[64] Experiential memory is *de re* memory of some particular. The requirement that there be a capacity for *conscious* memory is embedded in the second assumption. The requirement that there be an ability to remember inferential processes *de re* entails that the memory is experiential. It is *de re* of events in the inferential process.

What of the requirement of long-term memory? Long-term memory is a technical concept. It contrasts with short-term memory—which is, colloquially, *very* short-term. The various types of short-term memory last a few seconds, or less, and have automatic decay times. The memory required of a self in inference—in moral thinking or critical reasoning—is long term in this sense. At least, a being that had a reasonable lifespan and lacked a natural competence to hold together propositional inferences—or practical decisions or deliberative dialectic—longer than a couple of seconds would be rationally deficient in a way that would fall short of having the powers essential to being a person or self. (See note 102 below.)

Requiring that a capacity for meta-psychological autobiographical episodic memory in inference be *from the inside* entails requiring the memory to be capable of veridically preserving the perspective of the inference. This requirement follows from the second assumption. The relevant sort of *de re* thinking about one's inferential processes involves correctly remembering the mode and content of particular inferential steps through purely preservative memory. When meta-psychological autobiographical episodic memory relies on purely preservative memory in this way, the former type of memory preserves the perspective of the inference and represents it as it was when it occurred.

---

[63] *Purely preservative memory* is retention of mode and content from an earlier psychological state, retention that does not introduce new subject matter and is not part of a justification for the conclusion of an inference. For discussion of purely preservative memory, see my "Content Preservation," *The Philosophical Review*, CII, 4 (October 1993) 457–88; "Memory and Persons."

[64] See *ibid.* Recall that experiential memory need not be warranted empirically. It is not a type of inner *sense.* See note 37.

The two assumptions show that meta-psychological autobiographical episodic memory from the inside, or some short-term analog, is constitutive to selves. Such memory is clearly a type of consciousness of one's past self, plausibly in the sense of Kant and Locke.

I turn more briefly to a second way in which diachronic capacities are constitutive of being a self. I believe that both the special type of memory that I have been discussing and an analogous capacity for anticipation of the future are constitutive to selves, moral beings, and critical reasoners by virtue of their constitutively having *practical* reasoning. Practical reasoning has both forward-looking and backward-looking diachronic dimensions. Intentions have the representational function of being fulfilled by the intender. When they are fulfilled through occurrent decisions to act, and by actions themselves, there is a match with the antecedent intention. Self-conscious selves can recognize these points. Looking forward from an intention, an individual is aware of his or her being the prospective executor of the intention. The individual can think anticipatorily of him- or herself, in the first-person way, as preserving the intention and as realizing it, through a decision to initiate action and through the action.[68] Remembering backward from a decision to initiate action, or from an action, a self-conscious individual can recognize moving on the antecedent intention, and satisfying or failing to satisfy the intention. These diachronic capacities are integral to the self-understanding that is constitutive of being a self and of being subject to moral norms and norms of critical practical reason. They are consciousness of oneself as one will be in the future, as well as consciousness of oneself as one was in the past.

An argument similar to the one given for propositional inference shows that requirements on having the reflexive powers with regard to decision making that are required of selves requires meta-psychological autobiographical episodic memory from the inside—and an analogous *de re* capacity to anticipate particular representational events in one's future.

---

[68] Framing possibilities for action is an aspect of selves' making choices. Selves have a probably constitutive capacity to think of themselves not only diachronically, but modally. Locke writes of selves as being able to be conscious of themselves in different times *and* places. Cf. *An Essay Concerning Human Understanding*, 2·27:9. He may have had in mind the relatively primitive freedom of imagination to think of different spatial possibilities in action.

A third source of the requirement that selves have self-conscious diachronic capacities lies in the dialectical reasoning integral to deliberation. A condition on being subject to norms of critical reason and morality is to be capable of dialectical self-criticism.[66] A person or self must be able to stand behind a belief or decision after reflecting on it, to answer self-questioning, and to effect change of mind after self-criticism. Such dialectic requires an ability to think *de re* of one's thoughts, while being committed to them and while criticizing past or prospective commitments. In critically evaluating a present commitment, one must be able to conceive oneself as possibly giving it up. In changing one's mind, one must be able to conceive oneself as having given up an earlier commitment, knowing why. Recognizing actual or possible differences in one's point of view, retrospectively or prospectively, is part of being a *critical* reasoner. Diachronic forms of self-consciousness are part of recognizing what accountability is—a capacity to support or change one's position through rethinking or new information. Such capacities to recognize differences in one's own point of view, as they unfold in dialectical reasoning, are constitutive to being a self.

The diachronic use of self-understanding that is partly constitutive of apperceptive rationally accessible points of view combines *episodic* memory with *purely preservative* memory. Episodic memory gives one diachronic *de re* access, allowing states other than immediately present ones to have buck-stopping status and to be part of the thinking practice that one implements reasons within. Purely preservative memory gives one diachronic access to the relevant psychological elements on their own terms—also a central requirement on their entering into critical reasoning and moral accountability.

To elaborate this latter point: purely preservative memory retains the mode *type* and representational content of psychological state instances. Analogously, for anticipation of the mode and representational content of one's future point of view. Mode and content are the aspects of psychological state instances that are relevant to application of norms of critical reason and morality. If diachronic self-consciousness could not preserve the content and inferential order of past psychological events, one could not review one's inferences and commitments on their own terms. If diachronic self-consciousness could not anticipate the mode and content of instances of one's future points of view, one could not control one's future acts. Reflective review and anticipatory control are basic functions of a self.

---

[66] Dialectical reasoning is a special case of agency. So the points in this paragraph specialize the points in the previous paragraph. I distinguish practical agency and agency involved in theoretical reasoning. Dialectical reasoning can be either practical or theoretical. The agency involved in reasoning to intentions or to act-initiating decisions is practical.

I believe that the foregoing considerations provide a deeper characterization and motivation of the relation between selfhood, or psychological personality, and diachronic self-consciousness than we find in Locke or Kant.

V

I conclude by summarizing features of the relevant self-understanding. These features will ground discussion of its epistemic warrant. That topic will dominate the third lecture.

The self-understanding must be *meta-psychological.* The norms that I have discussed demand understanding the norms and applying them to one's point of view. To understand application of the norms, it is not enough to have *competence understanding*—the capacity to *use* concepts. One must be able to think about propositional attitudes, and their representational contents, as such. This is meta-psychological understanding.

Second, the meta-representational understanding must be *phenomenally conscious* and *rational-access conscious.* There must be a what-it-is-like quality to the understanding. And the lower-level, "understood" aspects of the point of view must be occurrently accessed by the individual's rational powers. The requirement that a self be phenomenally conscious, some of the time, is so basic that it is hard to motivate. Except in cases of understanding sensations and other states constitutively associated with consciousness, phenomenal consciousness is less important epistemically than in marking off the psychological kind, *self,* that underlies our deepest normative valuations. Those states that are constitutively associated with phenomenal consciousness (pleasure, pain) ground many primary moral valuations. Beings that lack a natural capacity for phenomenal consciousness do not have the moral claim on us that phenomenally conscious beings do. The requirement that a self be able to make accessible to conscious rational control some lower-level aspects of its point of view is just the product of combining the requirement of phenomenal consciousness with the role of self-understanding in applying the norms to one's own case.

Third, the understanding must exercise *self-conscious* meta-representation. A self must have a first-person self-concept and must apply it consciously. It must recognize aspects of its point of view as its own. In thinking *I believe that Brahms is greater than Chabrier,* I not only refer to myself with *I*; I take the belief that I represent to be mine. This capacity to represent elements of one's point of view as one's own is, of course, basic to the self-evaluation that grounds being subject to moral norms and norms of critical reason.

An aspect of this self-consciousness is that both the lower-level psychological element and the meta-representational element that constitutes the self-understanding are reflexively *understood*. One is accountable for the meta-representational self-understanding as well as for the understanding's subject matter, the lower-level psychological element.

Fourth, the understanding must include past and future as well as present elements in one's point of view. I have highlighted this constitutive requirement because I believe that it has been less well understood in the philosophical tradition than some of the others.

Fifth, the self-understanding must be *systemic*. This requirement underlies the fourth one. The self-understanding represents not just psychological states and their representational contents. It represents relations among the attitudes and the contents. Reason is essentially relational. An understanding that applies the concept reason must specify relevant relations. Moral evaluation of acts must relate them to beliefs and intentions. Even where moral norms evaluate attitudes, they must do so *relative* to information accessible to the individual.

Sixth, the self-understanding must be *specific*—from the inside. To ground norms of morality and critical reason, the self-understanding must specify psychological states in ways that preserve the point of view to be evaluated. The specifications must enable one to know what the point of view is on its own terms. Since the norms concern the perspective of the individual, an understanding that did not preserve the mode and content of the evaluated psychological elements would be too detached to render the norms applicable. Of course, much is forgotten. But an individual's having a *capacity* to preserve the mode and representational content of evaluated attitudes is constitutive to applicability of the norms.

Seventh, the relevant self-understanding must be *immediate*, non-inferential, in the apperceptive rationally accessible point of view.[67] Access to self-understood psychological elements is not via individual-level inductive inference. Psychological elements immediately accessible to self-understanding are starting points for expression of an individual's moral being or critically reasonable self. They have buck-stopping status for evaluation under the relevant norms. Psychological elements that are not immediately accessible cannot ground applicability of the norms of morality and critical reason. The starting points of, and transitions in, some inference must be capable of noninferential self-understanding.

---

[67] One can work out inductively things about one's point of view from inside it, as long as one starts one's inferences from a self-understanding that is immediate and from the inside. Wisdom about oneself lies in that direction.

There is an existence proof of individual-level noninferential elements of self-understanding. It is provided by Descartes' pure and impure *cogito* judgments. The structure of these judgments shows clearly that they are not inferential. When I judge I am thinking, or I [hereby] judge that Brahms is greater than Chabrier, or I intend to take the metro, the judgment is reflexive in a way that precludes individual-level inference. The judgment about the psychological state has, as a constituent part, the psychological state instance that the judgment is about. Judgment and subject matter are part of a single thought.

Although *cogito* cases do not involve any individual-level inference, unconscious psychological causal processes connect some relevant types of self-understanding with their psychological subject matters. For example, all self-understanding that includes meta-psychological autobiographical episodic memory relies on sub-individual psychological transitions. Whether these are propositional inferences is an empirical matter. I think that they probably are not, but I rest nothing on this belief. What is important is that these memories are not results of *individual-level* inferences. *Within the apperceptive rationally accessible point of view,* some of them are immediate. Any "easy" inferences to self-understood states in a rationally accessible apperceptive point of view must be open to noninferential self-understanding of the inferences' starting points and transitions. If an individual comes inductively to a view about a state that is not immediately or easily accessible, and relies entirely on the induction for self-understanding, the individual approaches one aspect of his or her point of view from the perspective of another. The understanding is not from the inside. For it does not preserve that point of view.

Psychological elements can have buck-stopping status only if they can be self-understood from the inside. They must be capable of being self-understood *in* the relevant psychological acts, states, or processes. These points apply to autobiographical memories. To self-understand past states in ways that give those past states buck-stopping status, some of the memories must be from the inside. Hence those states cannot be understood entirely on the basis of induction, within the apperceptive rationally accessible point of view.

An eighth feature of the self-understanding is its *de re particularity*. This particularity has several aspects. Recall that *de re* representation is representation that picks out a particular via a noninferential, not purely attributional or descriptive competence.

The understanding is *de re* in four ways. First, the application of the self-concept is *de re*. Application of the self-concept is not purely attributional or descriptive. It is effected through a rule that infallibly effects reference through the mere occurrent *use* of the concept.

Second, one's understanding of the *representational content* of one's psychological states is *de re*. Self-understanding uses canonical representation of representational content. Such representation uses the representational content in the very representation of it. When one understands one's belief that Federer will win a match, canonical representation of the belief's representational content (that Federer will win a match) refers to the representational content, and also uses that representational content. Such canonical representation is *de re*.

Third, self-understanding of elements in one's point of view specifies them by applying a temporal index. This indexing is *de re*. One singles out the time nondescriptively and indexically, roughly as *now* or *this time*. In self-understanding that uses episodic memory, there are two *de re* temporal indexes: one for time of the self-understanding, the other for the past time of the understood psychological element.

Fourth, the self-understanding is *de re* with respect to particular state instances, acts, or occurrences. One understands instances of psychological types, not just repeatables. In singling out instances, one uses concepts—the canonical name of the representational content, the self-concept, temporal indexicals, and concepts for psychological modes (like believes, intends, perceives as of). But the use is not purely conceptual. One also occurrently *applies* the concepts to yield *de re* understanding of psychological particulars. In thinking I am wondering if Israel will make peace, I refer to an occurrence of wondering. In thinking, noninferentially, I believe that Rembrandt is deep, I represent a standing instance of the belief.

It is constitutive to the role of self-understanding in grounding applicability of these norms, and in constraining their application, that it apply *reflexively* to the particular point of view and the particular psychological states of the understander, not just the same type of point of view. The individual must understand applicability of the norms to elements in his or her *own* point of view, *understood as such*.

I have been discussing *representational* aspects of the self-understanding that helps ground selfhood, ground the applicability of moral norms and norms of critical reason, and constrain application of these norms. It is meta-representational; conscious; self-conscious; diachronic as well as synchronic; systemic; specific; immediate, or noninferential, at the proprietary individual level of the psychology; and, in multiple respects, *de re*. Next lecture, I explore the *epistemic* status of the self-understanding—connecting representational structure and representational competence marked by that structure with their epistemic credentials.

SELF AND SELF-UNDERSTANDING

LECTURE III: SELF-UNDERSTANDING

I have taken selves to be constitutively self-conscious critical reasoners, and arguably constitutively subject to moral norms. I isolated a type of self-understanding required for applicability of these norms. The norms require an individual to be able to understand some of his or her psychological states from inside having those states.

Norms of critical reason and morality hinge on elements in an apperceptive rationally accessible point of view. Such a point of view is that part of a point of view whose elements are immediately or easily, presently, consciously accessible to self-attribution. Self-understanding helps give such elements *buck-stopping status* with respect to the two norms. Having buck-stopping status is having most fundamental status in explanations of the applicability and applications of a norm.

I

I want to understand the *epistemic warrant* for this self-understanding. To explain any epistemic warrant, one must explain what makes a warranted state an objectively good route to truth, given an individual's background information and competencies. Being a good route implies reliable veridicality in normal conditions. Normal conditions are conditions by reference to which the state's representational nature is constitutively explained.[88]

A warrant must also "make sense of" or rationalize the individual's point of view. This condition is hard to make precise. It is grounded in the idea that warrant is a norm for psychological competence. There must be something *about the nature of the individual's capacities or point of view* that helps explain the individual's psychological states' being a good route to truth. Brute, "accidental" reliability does not suffice for warrant.

It would be too much to require an individual to be able to explain the warrant. Animals, children, and even adult persons can be warranted in perceptual or other beliefs, and yet be unable to explain why. On the other hand, correct explanation, even if the individual cannot give one, must center on the individual's point of view. For what is being made sense of or rationalized, under epistemic norms, is the individual's competence and perspective.

---

[88] For an argument that normal conditions, in this sense, are the relevant conditions in which the reliability of the underlying warrant is to be considered, see my "Perceptual Entitlement." This account shows how the condition of reliability, and warrant for one's beliefs, can remain in place even in conditions—such as brain-in-vat conditions—in which one's beliefs are highly unreliable. Although *de facto* reliability—for example, in abnormal conditions—is not constitutive of being warranted in the empirical case, *de facto* reliability in one's access to some of one's own mental states is constitutive of being a self. Individuals that lose all reliability in judgments about any of their mental states, say through pathology, lose their selfhood.

Epistemic warrant for a belief is sufficient for knowledge, if the belief is true and if its relation to truth is in certain ways not abnormal or adventitious. Warrant concerns the contribution to goodness of route to truth by the *individual's* competencies. Failures of knowledge do not reflect specifically on the individual or the individual's psychology if the individual's states are warranted. In this way, warranted but false beliefs are like good acts with bad consequences. I call a false but well-functioning and warranted propositional attitude, warranted in way *W*, an attitude that incurs *brute error* relative to warrant *W*.[69]

Many epistemic warrants are compatible with brute error. Nearly all *empirical* warrants are compatible.[70] Perceptual beliefs can be warranted but mistaken because of an illusion that one lacks ground to reject. Reasonable inductive empirical explanations can be mistaken. Even most empirical *knowledge* rests on warrants that are compatible with error.

The warrants for self-understanding and self-knowledge that we are exploring show a different pattern. The warrants yield *immunity to brute error*: if the psychological state is warranted, it is veridical. Before developing this point, I shall discuss it intuitively.

Many psychological states are not easily accessible to self-knowledge. Then a belief's being warranted is compatible with brute error. Knowing one's deeper motivations, or character, or the unconscious states and transitions that populate the mind, is hard. Error can derive from the sheer difficulty of the psychological subject matter. Often, mistaken self-ascription is unwarranted. But warranted error can certainly also occur.

By contrast, in many types of cases, if a mature person is wrong about his or her beliefs, intentions, percepts, or sensations, it is common to infer that there is something wrong with the person. Errors reflect on the use or health of the person's cognitive equipment. If a person were wrong about his or her *belief* about whether he or she has a sibling, it would be normal to infer that the person was self-deceived, or had allowed emotion to cloud judgment, or was afflicted with some pathology. It is easier to imagine cases in which one has a warranted mistaken belief about whether one *has* a sibling. I am evoking the oddity of a warranted but mistaken belief about whether one *believes* that one has a sibling. Here error commonly reflects on how one formed one's meta-belief. Often, a mistake entails malfunction or other nonstandard belief formation, which undermines warrant.

---

[69] I introduce the notion of brute error in "Individualism and Self-Knowledge," this JOURNAL, LXXXV, 11 (November 1988) 649–63; cf. also "Perceptual Entitlement," section II.

[70] Exceptions are knowledge of one's sensations and perceptions. These are empirical because their warrants depend for their force on *having* of the sensation or perception. I later discuss why at least paradigm cases of these sorts of knowledge are immune to brute error.

These issues are delicate. The *natures* of *most* psychological states that are accessible to relevant self-understanding do *not* guarantee that they are accessible.[71] Being a belief that one has a sibling does *not* insure susceptibility to self-understanding that is immune to brute error. But I shall argue that, for selves, there are *necessarily some* cases of immunity to brute error, and that such cases ground applicability of norms of critical reason and morality.

I cannot emphasize too strongly that except in *cogito* cases, self-attributions that are warranted in the relevant way *cannot* be determined by the contents and modes of the attributed states. Evidence as to whether one has a sibling could be mixed. One could be unsure of the nature of one's own reaction to the evidence: do I believe that I have a sibling? Using behavioral evidence about oneself, one could be warranted in concluding that one does believe one has a sibling. But one could lack that belief. One could be puzzled, or have a leaning that falls short of belief. Then, the warrant for the mistaken belief that one believes that one has a sibling would not be the warrant that I am reflecting upon. The warrant would not be immune to brute error. Except in *cogito* cases, one cannot generalize apriori about *which* of one's self-attributions can be warranted in the special way. What is apriori determinable, I think, is that *some* psychological states in selves are susceptible to a type of self-understanding that is immune to brute error, relative to the relevant warrant.

Intuitively, we sometimes do understand our own psychological states in a specially warranted way. This intuitive point is supported by theoretical considerations. Let us reflect on the relation between immunity to brute error for the relevant self-understanding and the buck-stopping status of self-understood psychological states in critical reason and morality. Immunity of the self-understanding to brute error is constitutive to having an apperceptive core self. A type of self-understanding that is *liable* to brute error cannot, I think, ground the buck-stopping status of attitudes in applications of these norms. If an individual's self-understanding of a psychological element can be warranted but mistaken, the self-understanding is too detached from that element to give that element buck-stopping status. That is, if one can understand a psychological element only in a way that is liable to brute error, that element cannot be a basic starting point for assessing the individual's contribution to full unmitigated accountability under the norms of critical reason and morality.

---

[71] Exceptions are pure *cogito* cases and some knowledge of conscious sensations.

Being subject to brute error means that in reviewing one's psychological state or event, one is not fully accountable for getting that state or event right. Error is compatible with one's best use of one's competence for self-understanding. Such reviews do not constitute exercise of a self-understood point of view that one has cognitive control over. They are not exercises of a cognitively unified perspective. The norms of critical reason and morality take the basis for their evaluations to be psychological states that are understood from inside the perspective of those states—preserving that perspective meta-cognitively in the individual's self-review. The norms require the individual to be able to understand what he or she does or undergoes *in* the doing or undergoing, from the inside. If one could not take up the same perspective that those acts or endurings had from the inside—incorporating it into the meta-representational review, by preserving it or making it part of a single meta-representational perspective—one would always be in the position of a privileged outsider. Psychological elements self-understood in ways that allow for brute error lack buck-stopping status in persons and selves.

The idea is that, ultimately as a critical reasoner and as a moral being, an individual's self-understanding must inform, or be capable of informing, some of the individual's cognitions, reasoning, decisions, and other states central to being a self. The self-understanding must inform them in a way that is part of their exercise or part of their being experienced. Working out an inductive, empirical understanding of one's reasoning and decisions, or relying on some other sort of understanding of them that is subject to brute error, is not the kind of self-understanding that forms a basic starting point for critical reasoning or morally responsible decisions. The self-understanding must be from the inside—from a meta-perspective that preserves and is informed by—the psychological states and acts that one is most basically and unmitigatedly responsible for as a critical reasoner and moral being. One must be able to self-understand from the inside those states, acts, and events that are fundamental to being a self.[72]

---

[72] I make a version of this point in "Our Entitlement to Self-Knowledge," *Proceedings of the Aristotelian Society*, XCVI (1996). 91–116; cf. 109ff. See also the discussion of (CR) below

There are certainly types of self-understanding that rest on warrants that leave one vulnerable to brute error. Immunity to brute error is not an honorific status. It is not invulnerability to all error. It is a normative standing that hinges on possible sources of error, relative to epistemic norms. The deepest wisdom in self-understanding—a result of learning about oneself—*is* vulnerable to brute error. I claim only that the self-understanding required by norms of morality and critical reason—and required for there to *be* selves—is immune to brute error: being warranted is incompatible with error. This is a less comprehensive, less insightful self-understanding than wisdom about oneself. But it is the sort that is constitutive to being a self, and to the buck-stopping status of self-understood states in the applicability of norms of critical reason and morality. From here on, I *capitalize* ('Self-Understanding', 'Self-Knowledge') to indicate the self-understanding and self-knowledge that is my focus. *Sans* capitalization, the terms express generic notions that apply to all types.

II

In the last quarter century, self-knowledge has come in for a lot of discussion. I will not canvass the literature. I ignore views that assimilate Self-Knowledge to inner perceptual belief or to empirical hypotheses. I and others have criticized such views elsewhere.[73] I also ignore views that deflate Self-Knowledge, holding that it is trivial or nonsubstantial. I will, however, say a little to distinguish my view from others.

The psychological competencies that engender Self-Knowledge are not in general infallible. Pure *cogito* thoughts *are* infallible. If one thinks I hereby in this very act think [entertain] the thought that *Figaro* is the greatest opera, that thought about my thinking cannot

---

One can *get to* a veridical self-understanding that replicates the perspective of a psychological state by inductive inference. If the self-understanding relies on the inference, it would still be subject to brute error, as all induction is. But it would not have buck-stopping status. It could not ground the applicability of norms of critical reason and morality. Only the starting points of the inference could have buck-stopping status, assuming that they are susceptible to the right sort of self-understanding. The starting points of the inference could not have buck-stopping status and be subject to brute error, if they are to be *from inside* the perspective of the understood psychological states, in the relevant sense.

[73] Sydney Shoemaker, "Self-Reference and Self-Awareness," this JOURNAL, LXV, 19 (Oct. 3, 1968). 555–67; Shoemaker, *The First-Person Perspective and Other Essays* (New York: Cambridge, 1996), my "Individualism and Self-Knowledge" and "Our Entitlement to Self-Knowledge"; Richard Moran, *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton: University Press, 2001), Akeel Bilgrami, *Self-Knowledge and Resentment* (Cambridge: Harvard, 2006), chapter 1; and Lucy O'Brien, *Self-Knowing Agents* (New York. Oxford, 2007), chapter 3

be mistaken.[74] But pure *cogito* cases comprise only a tiny portion of Self-Knowledge. Immunity to brute error in Self-Knowledge is not immunity to all error. Noninferential self-attributions of the relevant types can be mistaken because of unwarranted interference or pathology. The epistemic specialness takes subtler forms than infallibility.

The competencies that yield Self-Knowledge are also not to be understood in terms of self-intimation. I doubt that *any* type of psychological state is such that if an individual is in it, it is impossible for the individual not to believe that he or she is in it.

Some traditional accounts of self-knowledge explicate its epistemic specialness in terms of the nature of the subject matter. Conscious mental elements are supposed to render judgments about them epistemically special because of the elements' very nature. Thoughts were said to be transparent, self-evident, unmistakable, or self-intimating by their very nature—as if they glowed in the dark. Such accounts are neither explanatory nor accurate about the psychological or epistemic phenomenon.

Some recent accounts have made claims—rhetorically different, but I think ultimately similar—that it is of the nature of 'genuine', 'full-blown' intentionality that mental states be known or normally knowable by any individual who has them.[75] Animals and very young children are attributed a lesser type of intentionality, or none at all. I think that such views are hopeless. Their conceptions of "genuineness" in intentionality lack serious defense. The accounts do not accord with scientific explanations that attribute perceptual, intermodal, and propositional intentional (representational) states to animals and very young children. Many or all of these individuals lack self-knowledge. They have intentional, representational states in the most literal sense of these terms. No argument has shown that their psychological states are any less genuinely representational than those of human adults. There is no difference in kind or degree in their *representationality*. Self-Knowledge, and self-knowledge, cannot be explained in terms of the natures of first-order representational states. There is nothing in their natures that insures that they are self-known or self-knowable.

---

[74] See my "Individualism and Self-Knowledge"; Andreas Kemmerling, "Eine reflexive Deutung des Cogito," in Konrad Cramer et al., eds., *Theorie der Subjektivitat* (Frankfurt, Germany: Suhrkamp, 1987); Kemmerling, *Ideen des Ichs: Studien zu Descartes' Philosophie* (Frankfurt, Germany: Suhrkamp, 1996) The original source is, of course, Descartes, *Meditations on First Philosophy*, II.

[75] Donald Davidson, "Rational Animals" (1982), "The Second Person" (1992), and "The Emergence of Thought" (1997), all in *Subjective, Intersubjective, Objective* (New York: Oxford, 2001); Bilgrami, *op. cit.*, cf. pp. 160ff., 178–82, 205.

The natural phylogenetic and developmental order is to have representational states before being able to know them. This natural order is the order of language acquisition. Even if higher-order representational concepts are innate in human beings, which I doubt, no one has given good reason to think that there is a constitutive connection between an individual's having representational states, even having propositional attitudes, and that individual's being in a position to know what those states are.

Another claim is that the subject matters of self-knowledge constitute *reasons* for the knowledge.[76] Both lay-people and philosophers think about reasons in disparate ways. I make just a few points here.

Reasons are necessarily representational. They are considerations that are good routes to truth, in the case of theoretical reasons, or good routes to goodness or aptness of acts, in the case of practical reasons. A thunderbolt or a sea otter cannot be a reason. Only some propositional representational content, with some associated mode—for example, judgment—about the thunderbolt or sea otter can be a reason. Nature *grounds* reasons.[77] It provides entities and situations representation of which can form a reason. But reasons are constitutively propositional and representational. Nature is not a text and is not made up of propositions or reasons. Since reasons are necessarily *representational*, sensations cannot be reasons unless they are representational. I believe that some sensations are not representational. In such cases, only recognition that one has the sensation can be a reason.

Reasons are necessarily *propositional*. Reasons are potential steps in arguments and explanations that, in effect, show why one should believe or intend something.[78] They answer potential why questions. As a matter of logical grammar, nonpropositional—and nonrepresentational—entities cannot play these roles. Nonpropositional entities, such as perceptions, can contribute to an attitude's being warranted. But reasons are necessarily propositional contents taken with their modes. They are basic units in reasoning. Some Self-Knowledge is of sensations and perceptions, which are not propositional. These entities *could not* be reasons for Self-Knowledge. But it can be shown independently of this point that what one knows in Self-Knowledge is not a reason that supports the knowledge. Let us focus on knowledge of one's own propositional attitudes.

---

[76] Christopher Peacocke, "Conscious Attitudes, Attention, and Self-Knowledge," in *Knowing Our Own Minds* (New York, Oxford, 1998).

[77] A major source of the confusion lies in not distinguishing support in reasoning and argument—reasons—from support of other kinds, including physical evidence For a good account of such distinctions, see Robert Audi, *The Architecture of Reason: The Structure and Substance of Rationality* (New York, Oxford, 2001).

[78] I say 'in effect' because the reason need not respond to an actual question, and because a reason is commonly not a meta-level consideration that makes reference to shoulds, beliefs, intentions, or representational contents. Reasons can be about object-level matters. They are reasons by virtue of their potential use in supporting and explaining the "why" of conclusions.

Are first-order beliefs reasons to judge that one has those beliefs? They cannot be. A reason, theoretical or practical, must be about the subject matter of the attitude for which it is a reason, or about some relevantly related subject matter. The reason's representational content must bear on, and in some way count in favor of, the truth or goodness of the conclusion for which it is a reason. First-order attitudes are not about the subject matter of second-order attitudes. They commonly bear no relevant relation to it. They *constitute* the subject matter, but they are not reasons about the subject matter that support the truth or goodness of the higher-order attitudes. If one's first-order attitudes were used as reasons for judgments that one had them, one's reasoning would consist of *non sequiturs*.

Most Self-Knowledge is not supported by reasons. The few exceptions are self-evident *cogito*-like cases. Not being supported by reasons does not make Self-Knowledge any less epistemically warranted or substantial. I shall elaborate this point shortly.

An idea with some currency is to explicate Self-Knowledge by reference to knowledge of one's intentional physical agency.[79] I myself emphasize Self-Knowledge's role in the applicability of norms of morality and critical reason. These norms centrally concern agency. It is certainly true that *some* of what is known in Self-Knowledge comprises mental acts. Persons and selves—the subject matters of Self-Knowledge—are constitutively agents.

The approaches that I have in mind, however, highlight agency in a more specific way. They model Self-Knowledge—or some types of Self-Knowledge—on an individual's knowledge of his own intentional *bodily* action. Knowing one's thoughts and attitudes is taken to be structurally and epistemically like knowing that one is raising one's arm.

I cite two grounds for not following this line. One is that doing so cannot account for Self-Knowledge's buck-stopping role in morality and critical reason. It is not one's bodily actions, but their psychological antecedents that ground these norms. Assessing full, unmitigated accountability for bodily actions depends on assessing their conforming to potentially Self-Understood cognitions, intentions, decisions, and intentional initiations that produce them. A corollary is that warrants for beliefs about one's bodily actions do not yield immunity to brute error.[80] The warrants I seek *do* yield immunity.

---

[79] See Moran, *op. cit*; Matthew Soteriou, "Mental Action and the Epistemology of Mind," *Noûs*, xxxxix, 1 (March 2005): 83–105; Bilgrami, *op. cit.*; O'Brien, *op. cit.*; Peacocke, "Mental Action and Self-Awareness (I)," in Jonathan D. Cohen and Brian P. McLaughlin, eds , *Contemporary Debates in the Philosophy of Mind* (Malden, MA Blackwell, 2007), pp 358–76; Peacocke "Mental Action and Self-Awareness (II): Epistemology," in O'Brien and Soteriou, eds , *Mental Actions* (New York: Oxford, 2009), pp. 192–214. Some of these approaches are not strictly incompatible with what I say here, but all highlight agency in self-knowledge more than I do.

A second ground for not modeling Self-Knowledge on knowledge of bodily action is that a lot of Self-Knowledge is not of acts. Selves *are* constitutively agents. Thinking morally or critically requires a self-consciousness and control that insure that the thinking is agential. Critical reasoning requires judgment, an act. Persons and selves constitutively make decisions in the light of values, again acts. All these points bear on constitutive aspects of being a person or self. But persons and selves, *we*, are not *just* agents. We not only act. We undergo and endure. Knowledge of passive elements in one's mind is virtually as important to norms of morality and critical reason as knowledge of active elements.

Some Self-Knowledge centers on sensations and perceptions that we do not bring upon ourselves. Some Self-Known propositional thoughts simply occur to one. Most emotions are not acts. Even many propositional commitments—all standing commitments (standing beliefs or fears) and many occurrent commitments (many occurrent perceptual beliefs)—are not acts. Many do not even derive from acts. Many are acquired through perception and are first stored automatically.[81] Sometimes in Self-Knowledge one reaffirms such states in judgments—which *are* acts. But often in Self-Knowledge one simply forms a belief about what beliefs one has long had. The truth and warrant involved in such meta-beliefs do not hinge on reaffirming the standing commitments that did not arise through any agency. Such non-agential states and occurrences are susceptible to Self-Knowledge.

---

[80] See discussion in Lecture II, section III of the buck-stopping status of the psychological antecedents of bodily actions. The immunity to brute error of Self-Knowledge of the psychological antecedents derives from the same fact that gives the psychological antecedents buck-stopping status. Knowing the psychological antecedents depends only on good use of the competence for Self-Understanding. Knowing one's actions through knowing the psychological antecedents is subject to contingencies involved in realizing one's intentions. One can be fully warranted in one's belief about what one's action is, utilizing one's warranted belief about what one is deciding to do. But one can still be mistaken about the action because of occasion-dependent, post-initiation distortion of the bodily action. Then, the individual's contribution to accountability lies fundamentally in the Self-Understood, or Self-Understandable, psychological antecedents. Accountability for the bodily action commonly is mitigated. Mitigation tracks the individual's brute error about what the action is.

[81] Many philosophers emphasize consciousness as a constitutive feature of intentionality or even knowledge. There are, however, many unconscious representational and knowledgeable states. If consciousness makes a type of state more reliable, it is epistemically relevant. This is an empirical matter. There is probably unconscious self-knowledge that is analogous to Self-Knowledge in being epistemically "special." Any such self-knowledge is not part of the apperceptive core self.

Knowing one's pains, fears, and other aspects of vulnerability is crucial to one's moral status. Knowing one's perceptual perspective and other passive representational states is crucial to critical reasoning. Self-Understood passive elements in a psychology loom large among elements with buck-stopping status for both morality and critical reason. Indeed, Self-Understanding itself can be passive—the product of the triggering of standing competencies.

III

Let us turn from paths not taken to our main route. I noted earlier that most Self-Understanding and Self-Knowledge are not based on reasons, but are nonetheless warranted. I distinguish two types of epistemic warrant. A warrant that is not based on a reason is an *entitlement*. Such warrants need not be accessible, even in principle, to the individual whose thoughts or attitudes are warranted. A warrant that is based on a reason is a *justification*.

Apart from *cogito* cases, I think that the special epistemic warrant underlying Self-Knowledge is an *entitlement*, not a justification. The warranted individual need not have a reason for the warranted belief. On the other hand, the entitlement requires more from an individual than many other entitlements do. A young child can be entitled to perceptual beliefs, even though it has only the competence understanding required to have the beliefs. To be warranted in Self-Knowledge, an individual must be competent in employing meta-representational attitudes, and must think the content of the attitudes that are the subject matter of the Self-Knowledge, while specifying them in the canonical, that-clause-like way.

I believe that warrants for attitudes that constitute Self-Understanding and Self-Knowledge yield *immunity to brute error*. If an attitude *type* is associated with competencies that could ground a warrant W_{ibe} that yields immunity to brute error, any error in an *instance* of an attitude of that type must be explained in one of three ways. First, the attitude instance may be warranted only in another way W that does allow brute error. Second, the instance may derive from misuse of competencies that undermine warrant W_{ibe}. For example, bias, self-deception, or emotion might derail a normally warranted belief formation. Third, the individual might suffer some pathology that undermines W_{ibe}. A lesion might cut off judgment from normal access to intentions or beliefs. In such cases, an individual lacks the *relevant* epistemic warrant. Having an epistemic warrant that yields immunity to brute error guarantees *getting* *things right.* Being relevantly entitled to an attitude about one's psychological states insures that the attitude is true.

Entitlements to Self-Knowledge are not the only warrants that yield immunity to brute error. I shall reflect on three other classes of cases.

The first class includes three types of entitlement associated with first-order deductive inference. In discussing this class, I distinguish first-order inference from such inference *supplemented* by thinking the inference rules that warrant steps in the inference. Individuals can carry out inferences without having the meta-representational concepts, or the ability to generalize, necessary to think the inference rules that justify their inferential transitions. Even if they do have these abilities, the abilities may not be operative in a good deductive inference. Reasoners can be warranted in their reasoning if the reasoning is governed by—correctly explainable in terms of—following correct rules of inference, even if the reasoner lacks resources to think the meta-representational rules, consciously or unconsciously, or does not rely on such resources in carrying out the inference. The *steps* in the inference can be justifications, reasons, for later *steps*. The warrants for the transitions between steps can, however, be entitlements—warrants without reasons. Reasoners can be entitled to rely on the transitions as valid, even though they cannot, or do not, think the inference rules. Entitlement derives from competence with the logical constants in carrying out the reasoning.

The entitlement to make a deductively valid transition yields immunity to brute error in making the transition. One can make a mistake in such transitions. But if one is entitled to a transition in a deductive inference that relies on purely preservative memory, one cannot make a mistake in the reasoning transition.[22] Being entitled to a transition entails that the transition is free of error.

---

[22] Equivocation between "twin" concepts in "slow-switching cases" requires that this point be understood with some subtlety. I think that errors in reasoning that depend on reasoners' mixing up "twin" concepts that they have unawares rely on mistaken presumptions of equivalence between the twin concepts. In this respect they are like mistaken presuppositions or even mistaken tacit beliefs. The errors do not result from failures of anaphoric, purely preservative memory, since such memory is not relied upon in the inference. They occur when a reasoner relies on sources—such as substantive memory or perception—that introduce content into later premises in an argument *without* relying on purely anaphoric, preservative relations to the content of earlier premises. Nor do the errors result from failures of rational deductive competence. They involve what Luca Struble calls a brute error in exercising a coordination competence, not an error in a reasoning competence *per se*. Although the presumptions of sameness or equivalence of "twin" concepts are not tacit beliefs, since no state puts the two concepts together into a proposition, the presumptions have much the same function. It is clear that possible errors in inference that hinge on equivocation between

A second entitlement is involved in fulfilling the first. This is entitlement to rely on purely preservative memory. Purely preservative memory is anaphoric retention of representational content and mode type. This entitlement, too, yields immunity to brute error. One cannot be warranted in relying on purely preservative memory and make a mistake in the preservation. Memory errors in reasoning derive from allowing some other type of memory besides purely preservative memory to enter the reasoning, or from some malfunction in purely preservative memory.

A third entitlement is present in those deductive inferences that involve coming to some reason-supported conclusion. Rules for deductively valid transitions are not rules for using deductive inference to provide reasons.[33] The former rules constrain permissible reasons, but do not guarantee that premises that lead validly to a conclusion are reasons for the conclusion. The premise of a deduction could be unwarranted. Then a valid deductive transition from it would provide no reason for the conclusion.

An individual could make an inference between propositions that are in fact connected by a valid rule of inference, without relying on the logically valid connection. The individual might make the inferential connection by rote, for example. Such transitions are not really deductive inferences. They are surely not reason supporting by virtue of deduction. Such inferences are not correctly psychologically *explained* in terms of valid logical inference rules. Inferences that are correctly explained in terms of valid logical inference rules must

---

"twin" concepts are brute errors. They do not reflect on the well-functioning of *memory* or of *reasoning competence*. So the errors do not have their sources in memory or reasoning competence. I introduced slow-switching cases in "Individualism and Self-Knowledge." Paul Boghossian thinks that such cases cause problems for anti-individualism in accounting for self-knowledge and inference. See his "Content and Self-Knowledge," *Philosophical Topics*, XVII (1989): 5–26; and "Externalism and Inference," *Philosophical Issues*, II (1992): 11–28. I think that there are multiple grounds for finding his arguments unsound. I discuss these issues briefly in "Memory and Self-Knowledge," in Peter Ludlow and Norah Martin, eds., *Externalism and Self-Knowledge* (Stanford: CSLI Publications, 1998). There is further literature on this topic that I will not go into here. The points in the text here supplement my 1998 discussion. For a somewhat different view of reasoning in slow-switching cases, see Mikkel Gerken, "Equivocation and Warrant by Reasoning," *Australasian Journal of Philosophy*, LXXXIX, 3 (2011): 381–400. I regard Gerken's view as broadly congenial. It could be accommodated with minor revisions of the points made here.

[33] I am using Gilbert Harman's point that rules of logic and rules of reasoning are not identical. See Harman, *Thought* (Princeton: University Press, 1973). One can engage in valid, even sound, deductive inference even if the premises provide no reasons for the conclusions.

involve, and must be further explained by reference to, exercise of a competence with the logical constants—not a rote connection, or some other nonlogical connection.

For an inferential transition to provide *reason support* for a conclusion by way of a deduction, the transition must be correctly explainable in terms of deductive inference rules and competence with logical constants; the premises of the inference must be warranted; and the reasoning must be relevantly noncircular. For a deductive reason-supporting transition to be correctly explainable in terms of deductive principles, its premise must ground some rational explanation of why the conclusion is worthy of belief *for the individual*. The individual need not be able to give the explanation. But the individual's competence with logical constants must rationalize—ground a potential explanation of—why the conclusion is belief worthy for the inferrer. Reasons are answers to potential why questions. An individual's competence with logical constants must comprise some (non-meta-representational) competence understanding of the connection between warranted premise and inferred conclusion—a connection that is belief supporting.

Now, a point analogous to the point about entitlements to transitions as deductively valid applies to entitlements to transitions as reason transmitting or reason supporting. These entitlements again yield immunity to brute error. If a transition in an inference is made with the function of providing a reason—if the individual reasons to commitment to a conclusion that is purportedly supported by the premises, not merely deductively infers the conclusion from the premises—the individual is entitled to the transition step, even lacking an ability to think meta-representationally about it, if and only if the transition is in fact governed by (correctly explainable in terms of) correct principles of reason support.[84]

These entitlements in deductive inference and related ones in inductive inference are, I think, developmentally the first warrants that yield immunity to brute error.

I turn from this first class to a second class of warrants that yield immunity to brute error. This class is comprised of warrants to believe simple, self-evident truths on the basis of understanding them. If one understands I am now thinking, either snow is white or it is not the case that snow is white, or $2 + 2 = 4$ well enough to be justified in believing it through understanding it, being warranted insures not being mistaken in the belief.

---

[84] Rules for transitions in inductive inference are not well understood. But it is clear that if a step is warranted, it is not inductively fallacious, although it could lead to a false, warranted conclusion. Reliance on the transition rule, not the conclusion, is immune to brute error.

A third class of cases of immunity to brute error comprises entitlements to *believe* noninferentially *in simple cases* that one attitude, or content together with a mode type, is a reason for another. *Simple cases* are those in which the individual uses a lower-level reasoning transition—entitled in the way of the third example in the first class—as a basis for the meta-level judgment about the transition relation. This class is the meta-representational analog of entitlements to first-order reason-supporting transitions. I distinguish belief that one thought is a good reason for another from belief about the form of the inference, and from theoretical generalizations about inference. I am concerned here purely with simple, *de re*, evaluative beliefs *about* particular transitions: *P* is good [or bad] reason for *Q* [perhaps relative to background *B*], where the individual can and does reason correctly—*p* so *q*.[35] In simple cases, if one's noninferential understanding of the reason relation between particular attitudes is good enough to entitle one to the meta-belief that one attitude is a reason for another, the understanding is veridical: the reason-support relation holds. One can be warranted but mistaken in beliefs about reason relations. But such mistakes rest on inference in theorizing, or on complex cases. In simple cases, warranted judgments are true. Errors derive from irrationality or other failures of understanding that undermine warrant.

IV

The key to understanding warrants that yield immunity to brute error does not lie in necessity. There are warrants to believe necessary truths that do not yield immunity—warrants to believe empirical necessities. There are warrants that yield immunity that do not attach to necessary truths—*cogito* cases.

The key to understanding warrants that yield immunity to brute error does not lie in apriority. Some inductively based mathematical conjectures are, I think, apriori warranted but not immune to brute error. Not all warrants that yield immunity to brute error are apriori. Entitlements to beliefs about one's sensations depend for their warranting force on having the sensation. They are not apriori. Such entitlements can still yield immunity to brute error.

What grounds epistemic warrants that yield immunity to brute error? What distinguishes these three classes of competencies? What distinguishes other competencies that yield immunity to brute error? I think that the combination of four features grounds the immunity.

---

[35] I believe that the points apply to prima facie reasons as well as conclusive reasons.

First, all the relevant competencies are, or rest psychologically and epistemically on, one or another type of propositional understanding.

Second, all the relevant types of understanding, including but not limited to competence understanding, are constitutive of having certain explanatorily or normatively significant types of point of view. In order for the type of point of view to be possible at all, these types of understanding must, in their natural undamaged states, be reliable.

Third, fulfilling representational functions of the states that are, or rest on, these types of understanding either forges constitutive connections *within* the relevant type of point of view, or makes commitments that express an understanding of such constitutive connections.

Fourth, representational success in fulfilling representational functions does not depend on anything beyond exercising the understanding in a psychologically well-functioning way. In particular, being representationally successful (being veridical, preserving veridicality, and so on) on any given occasion does not depend on anything outside the psychology on that occasion, and does not depend on having any information beyond that involved in exercising, in a well-functioning way, the understanding that is constitutive to having the relevant type of point of view.

Immunity to brute error is derivable from these four features of a representational competence, together with this constitutive principle governing warrant:

> (E) A propositional state or occurrence is warranted, on a given occasion, if and only if it is the result, on that occasion, of the exercise of a representationally well-functioning propositional competence that provides a representationally reliable, epistemically good route to veridicality, allowing for the natural limitations of the competence with respect to its subject matter, and limitations of the information available to it.[86]

By the first of the distinctive features, the relevant psychological states or occurrences rest on a reliable propositional competence. By the third and fourth, veridicality on a given occasion does not depend on extra-psychological conditions or on specific information beyond the minimal information necessary to exercise the understanding. By the second, third, and fourth features, being representationally successful (being veridical, preserving veridicality, and so on), on a given occasion, depends only on exercising the understanding of connections within the point of view—the understanding constitutive to having the relevant type of point of view—in a well-functioning way. It follows that a propositional state or occurrence is warranted if it relies only on a well-functioning exercise of a relevant type of understanding. Any malfunction of a relevant competence undermines warrant for a state or occurrence that relies on the competence. By the fourth feature, any error in a state that relies on a relevant type of understanding derives from substandard functioning of the competence—and hence undermines warrant. So any *warranted* state or occurrence that relies only on the understanding is representationally successful. In other words, reliance on any of the relevant types of understanding is immune to brute error.

---

[86] For the framework that embeds (E), see my "Perceptual Entitlement." The point of the phrase 'allowing…' in (E) is to indicate that as long as the competence is reliably veridical, and functions representationally *well* in using its available information—in a way conducive to getting things right in normal circumstances—exercise of the competence can be warranted. I assume that natural limitations include fallibility. Warranted failure with respect to veridicality derives either from natural limitations—such as natural imperfections in perceptual resolution—or from using *well* information that is in fact misleading.

Let us return to the three classes of warrant that yield immunity to brute error with this grounding in mind.

Take the first class of cases. A reliable competence understanding that consists in making valid transitions in deductive inferences that rely on purely preservative memory is constitutive to any propositional point of view, starting with the simplest empirical ones. To have states with propositional structure, one must be competent to use such structure reliably in inference, including deductive inference. Successful exercises of this competence forge constitutive connections within a propositional point of view. If the competence is exercised in a well-functioning, unhindered way, the transitions are representationally successful. They preserve truth in virtue of form. An individual is entitled to rely on well-functioning exercises of the competence—on the transitions—as deductively valid. So well-functioning exercises of the relevant competence understanding are warranted only if transitions are deductively valid.[87]

The same points apply to the competence understanding in exercising purely preservative memory. Having a naturally reliable competence of this kind is constitutive to having any point of view capable of propositional inference, indeed any representational competence at all. The competence makes constitutive connections in a representational point of view. If the competence is exercised without malfunction or hindrance, it is representationally successful in preserving mode and content from earlier states. Individuals are entitled to rely on exercises of a reliable competence that is well-functioning and unhindered. So exercises to which individuals are entitled are representationally successful in preserving mode and content. So the entitlements yield immunity to brute error.

---

[87] The rules for deductive transition that meet this description are restricted to simple cases. See the definition of '*simple cases*' above. Norms that are not explanatory of an individual's actual reasoning, because they are too complex, are not relevant to the account

Similarly, for competence understanding that yields reason-supporting transitions in reasoning. A reliable understanding of this sort, applied in simple cases, is constitutive to forging structural connections in any propositional point of view. Almost trivially, individuals capable of reasoning are competent to make moves between propositional steps that, relative to their background information, are in fact reason supporting—as long as their faculties do not malfunction and as long as internal conditions do not hinder natural exercise of the competence. Exercise of the understanding unhindered by disease or bias yields reason-supporting transitions. Being entitled to such transitions guarantees representational success—reason-supporting transitions. So entitlement yields immunity to brute error.

The second class of warrants that yield immunity to brute error concerns epistemic starting points, not transitions. They are otherwise similar. Understanding simple logical and mathematical truths well enough to be warranted through understanding them constitutes an explanatorily and normatively significant type of point of view. That type of point of view is a natural psychological kind. Believing the relevant truths, on the basis of understanding them, is a reliable route to truth. In fact, such beliefs are rational starting points in reasoning to other truths. Any error is a failure of the sort of understanding whose successful exercise constitutes the type of point of view. Warrant for believing simple logical and mathematical truths lies in understanding them.[88] If one is warranted in believing them, one's belief is true. So the acceptance relative to the warrant is immune to brute error.

What of the third class of cases—meta-representational understanding of reason-support relations? Competence to make reliable judgments about reasons *as* reasons is distinctive of critical reason, certainly an explanatorily and normatively significant type of point of view. Recall that *simple cases* are, by definition, those in which the individual uses a warranted lower-level reasoning transition as a basis for the meta-level judgment about the transition relation. In simple cases, being warranted guarantees true judgments. For they function to forge certain constitutive structural connections within a point of view capable of critical reason, by recognizing constitutive structural connections within the lower-level propositional point of view. The meta-representational judgments are grounded in lower-level thinking of the reason relations *in* the meta-representational thinking about them. To get things right, the judgments need use no other information than that involved in the meta-representational understanding that conceptualizes, while relying on the competence with, the warranted lower-level reason-supporting transitions. Forging constitutive structural relations in a critically reasonable point of view just is correctly representing the warranting relations in the transitions in the lower-level reasoning that one evaluates. (As indicated in the first class of cases, the lower-level entitlements themselves yield immunity to brute error.) Any error in using the competence is a malfunction of the competence. Being warranted in using the normally reliable competence in simple cases requires that the competence function well on the occasion of use. So since any error in using the competence in simple cases is a failure in its function to establish constitutive connections within a critically reasonable point of view, any error in simple cases undermines warrant.

---

[88] Understanding self-evident logical truths is, I think, an abstraction from deductive transitions in reasoning. Understanding truths of *pure* logic is, I think, a further abstraction from instances of such principles. Understanding simple truths of arithmetic is, I think, an abstraction from counting. Understanding other simple mathematical truths, such as the simplest axioms of set theory (for example, the axiom of extensionality), is not an abstraction from particular cases but is part of a minimum explicational understanding of relevant concepts (here, the concept *set*). None of these truths is vacuous in the positivists' sense.

In understanding warrants that yield immunity to brute error, it is instructive to reflect on the contrast class. The most primitive warrant that allows brute error is entitlement to perceptual belief. Like all warrants, including those that yield immunity to brute error, perceptual entitlements are explicable only against a background of veridical states. But there are crucial differences in this background.

The first, less basic difference concerns relations between competencies underlying the warrants and having an explanatorily and epistemically significant type of point of view. An individual can have an empirical point of view, even one with warranted beliefs, and yet not be in a position to get any perceptual beliefs right.[80] How could such a thing happen?

[80] I think that something stronger applies. An individual can have perceptual beliefs, but no warranted ones. The nature of perceptual states does not require that they be reliably veridical in their normal content-determining environment. (Of course, most perceptual states *are* thus reliably veridical. See "Perceptual Entitlement" and note 68 above.) I have been misunderstood to hold, in that very article, that it is apriori and constitutively true that perceptual systems are reliable in normal circumstances—the circumstances in which their contents were formed. I do not hold this view, and consciously wrote around it in that article. An example of the misunderstanding is

An individual can inherit perceptual equipment that obtained representational content through patterns of interaction, including veridical applications, in a certain environment. Call this environment '*the normal environment*' for that equipment. (See notes 68 and 89.) Suppose that when given appropriate proximal stimulation, this equipment produces types of perceptual beliefs that would be reliably veridical in the normal environment. However, any individual can be placed in a non-normal environment in which, despite similar proximal stimulation, every perceptual belief lacks the kinds of distal antecedents that gave content to the individual's perceptual equipment. Then the individual has an empirical point of view. The individual's cognitive equipment works as well as it can, and reliably, relative to the warrant-determining normal environment. Yet nearly all the individual's perceptual beliefs, including warranted ones, are false.[90]

An analogous situation is not possible with types of understanding that ground immunity to brute error. Inability to exercise those types of understanding *veridically* would make having the relevant sort of point of view impossible. An individual that could not naturally make valid or reason-supporting deductive inferences would lack a propositional point of view. An individual that lacked an understanding that when naturally exercised yielded belief in simple logical or mathematical truths would lack a capacity to understand logic or mathematics.[91] An individual whose natural competence did not yield true beliefs about simple reason-support relations would lack critical reason.

A second, more fundamental difference between perceptual entitlements and entitlements for types of understanding that are immune to brute error helps explain the first. The difference is that well-functioning, unhindered exercise of the perceptual competence

---

Anthony Brueckner, "Content Externalism, Entitlement, and Reasons," in Sanford C. Goldberg, ed., *Internalism and Externalism in Semantics and Epistemology* (New York: Oxford, 2007), pp. 160–76. I had earlier been guilty of holding, without argument, that the nature of a perceptual state requires reliability. See "Our Entitlement to Self-Knowledge," p. 106n11. But I came to give up the view in the late 1990s. Obviously, if there were a good apriori argument for it, there would be a very direct and comprehensive argument against skepticism.

[90] Clearly, the same sort of point can be made even if the individual is located in its normal environment. Every perceptual belief results from a proximal stimulation with artificial, nonstandard distal antecedents. The deeper point is that constitutive explanation of the content of perceptual states in terms of *veridical* cases, which is I think apriori necessary, does not necessarily depend on the veridical cases' being the most common cases.

[91] Understanding logic and mathematics, beyond mere competence understanding, contrasts with reasoning logically or mathematically. It involves minimal explicational capacity. I think that such understanding is a stage of development that marks a significant type of point of view.

underlying perceptual-belief formation does not insure representational success. Success on particular occasions depends on nonpsychological, "brute" connections to distal matters, beyond the point of view, in the environment. The distal matters can vary so as to yield either success or failure, even while the psychological competence for forming perceptual beliefs functions well and in an unhindered way.[92]

Forming a given kind of perceptual belief depends only on antecedent psychological sets, psychological formation laws, and proximal stimulation.[93] Representational success depends partly, on each occasion, on distal matters—beyond the individual's psychology. Distal matters can vary while producing relevantly similar proximal stimulation. As a result, a reliable competence for forming perceptual belief that functions as well as it can in given circumstances—given the information available to it—can produce warranted but false belief—brute error. No analogous distal, subject-matter variation can occur with respect to well-functioning warranted exercises of the sorts of understanding that are immune to brute error. With such understanding, the constitutive connections—and the connections that representational success depends on—are entirely within the relevant type of point of view. Moreover, exercising the well-functioning understanding that is constitutive of a relevant type of point of view just *is* successfully traversing, preserving, or recognizing the relevant connections within the point of view. Getting things right does not depend on conditions outside the psychology and does not require information beyond that which is necessary to exercise the understanding. Well-functioning, unhindered, warranted reliance on the competence yields representational success on each particular occasion and in any environment. So well-functioning, warranted exercises are immune to brute error.[94]

---

[92] See my "Our Entitlement to Self-Knowledge"; "Perceptual Entitlement", "Disjunctivism and Perceptual Psychology," *Philosophical Topics*, xxxiii, 1 (Spring 2005): 1–78, especially sections vi–vii; "Disjunctivism Again," *Philosophical Explorations*, xiv, 1 (2011): 43–80, especially sections ii–iv.

[93] This is a statement of the Proximality Principle, which guides the science of perceptual psychology. See my "Disjunctivism and Perceptual Psychology," section iv; *Origins of Objectivity*, p. 385ff.; "Disjunctivism Again," sections ii–iii, v. Other issues in this paragraph are discussed in "Perceptual Entitlement."

[94] The fact that a reliable competence represents only internal psychological elements does not suffice to make the representational states that rely on it immune to brute error, relative to warrants for them. Reliable inductive reasoning from warranted premises about one's psychological states can yield warranted but mistaken beliefs. Representational competencies that are basic constitutive elements or basic constitutive connections in a point of view are immune to brute error. Such competencies include purely preservative memory, inferential transitions, using reasons, understanding basic logical and mathematical truths, recognizing simple reason-support relations as such, and the sorts of self-understanding that I will discuss.

Of course, default entitlements to perceptual belief are not the only entitlements that allow brute error. Similar points apply, however, to entitlements to empirical substantive memory and to anticipation of one's bodily action or external events. Mistakes can be warranted because they depend on non-normal relations to the environment on given occasions. One can exercise the reliable competence unhindered and be mistaken.

An interestingly different class of warrants that allow brute error comprises the products of various types of inductive reasoning. I have three types of induction in mind—empirical inductive reasoning about the environment, for example, inference to the best explanation from perceptual beliefs; inductive reasoning to conjectures in mathematics; and inductive reasoning about oneself. The conclusions of these types of reasoning can be warranted (justified), but mistaken. For example, in forming well-reasoned judgments about one's character, one can be mistaken. The induction might be based on behavioral evidence, or on types of self-knowledge that *can* be arrived at through Self-Understanding.

Inductive reasoning uses patterns found in a base case, or cases, to form judgments about a further case, or cases—relying on the likelihood that the pattern reapplies. Well-functioning, warranted use of induction never insures veridicality of the conclusion.[96] It is inherent in the method that even if the beliefs that form the induction base are true, the conclusion can be justified but mistaken.

In all these cases of capacities that are vulnerable to brute error, representational success—veridicality—depends, on each occasion, on more than optimal representational functioning of psychological competence, given the information necessary to exercise it at all. Not so, with the competencies that yield immunity to brute error. Although relying on them can yield failure of veridicality, such fault lies in themselves, not in a brute failure to match a subject matter beyond them.

V

By reflecting on other cases of warrants that yield immunity to brute error, I hoped to illuminate Self-Understanding. I now return to that topic. Self-Understanding is necessary to having an apperceptive rationally accessible point of view. Immunity to brute error in Self-Understanding follows the pattern established by other examples of immunity. To explicate entitlement to Self-Understanding and

---

[96] I exclude mathematical induction that occurs in Peano arithmetic. I think that warrant that derives from relying on such induction does yield immunity to brute error.

Self-Knowledge in detail, I must elaborate structural considerations set out at the end of Lecture II.

Self-Understanding and Self-Knowledge involve predication of a psychological element to oneself. I distinguish several aspects of the predication. These elements mark specific competencies that ground specific warrants.

Consider the predication in I believe that Mozart loved Haydn. I refer to myself with a singular application of the self-concept *I*, and I attributively apply the concept believe that Mozart loved Haydn to myself. This concept consists of three primary components—the verb-concept believe that indicates a psychological mode (belief), the present tense of the verb-concept, and the canonical singular concept that Mozart loved Haydn that denotes a representational content. I shall say that the full concept, believe that Mozart loved Haydn, *indicates* a *type* or *kind* of psychological element—belief that Mozart loved Haydn. It also thereby indicates the mode type, belief.

In veridical predications in Self-Understanding, the attribution establishes a *de re* relation to *instances* of those types—here, an instance of a belief that Mozart loved Haydn and an instance of the mode, belief. I shall say that, in such cases, the predicate concept that indicates the psychological type or kind *betokens* the instance of the type or kind. So the predicate concept *indicates* the psychological kind and *betokens* the instance of the kind. Both indication and betokening are types of representation.[96]

Predication in Self-Understanding involves attribution as well as indication and betokening. When attribution is veridical, the indicated *kind* is attributed, and the betokened *instance* is also attributed. These are different attribution relations, but I will call them both '*attribution*'. In Self-Understanding, kind and instance are attributed to the individual picked out by application of the singular self-concept. Although all these elements work together in the relevant predications, they are different aspects of the predication and correspond to different aspects of understanding. Thus there is *indication understanding*, *betokening understanding*, and *attribution understanding*.

Betokening is necessarily guided by indications of a kind—in veridical Self-Understanding, the indicated kind. Indication understanding is always a component of betokening understanding. I will not discuss indication understanding separately.

[96] Betokening is the predicative analog of syntactically singular *de re* reference. Not all predications involve betokening. The predication 'Some people love others' does not, if the quantifier is not associated with *de re* thoughts about particular instances of loving.

Betokening understanding is usefully thought of as a *vertical* relation. It is a relation between a higher-level understanding and a lower-level psychological instance that is understood. Both understanding and what is understood are elements in a point of view. The understanding establishes the sort of multi-tier point of view discussed earlier.

My main task is to explain wherein being relevantly warranted in betokening understanding entails *correct* betokening understanding. The explanation should account for the warrant's intuitively yielding immunity to brute error. I begin with Self-Understanding of *propositional* states and occurrences. Later I consider other types of Self-Understanding.

In tracking what follows, one must bear firmly in mind a distinction between lower-level competence understanding and meta-representational understanding. *Lower-level competence understanding* is just a competence to think with propositional contents. *Meta-representational understanding* is a competence to think about the exercise of the lower-level competence, or the modes and contents that are instantiated in the lower-level competence.

In the meta-representational betokening understanding integral to Self-Understanding, a propositional state *instance* or *occurrence* is betokened *de re*. The instance is betokened via its constitutive aspects—its representational content and its mode. In attributing each of these aspects, meta-representational betokening understanding makes essential use of lower-level competence understanding.

Recall from the previous lecture that in attributing representational content to one's propositional state and occurrence instances, one must exercise lower-level competence understanding in the represented thought content as one uses the canonical representation that names that content.[97] An under-recognized aspect of betokening understanding is that it involves an analogous phenomenon in use of the predicate for the *attitude mode* of the propositional content. Application of the concept believes in betokening understanding does not use a *represented* representational content. For believes does not indicate a representational content. But one can preserve in memory, with application of believes, the belief mode; and one can retain dispositions to implement reasons connected to any given belief. That is, in betokening understanding of an attitude instance, one exercises lower-level memory that preserves representational content—of which one has lower-level competence understanding—and preserves sensitivity to its being believed.[98] Further, in understanding the belief, one can prime dispositions to associate it with other attitudes that bear lower-level reason connections to it. In betokening understanding, one uses the lower-level preservation through memory of the attitude mode, along with the content, in picking it out, *de re*. One retains in memory the belief mode as one represents it. I will reflect, first, on aspects of this meta-representational competence and, then, on epistemic warrants for its uses.

---

[97] For more on the canonical representation of representational content, see my "Individualism and Self-Knowledge", "Postscript: Frege and the Hierarchy," in *Truth, Thought, Reason. Essays on Frege* (New York: Oxford, 2005); "Postscript: 'Belief *De Re*'," in *Foundations of Mind*; and "Five Theses on *De Re* States and Attitudes," section v.

Purely preservative memory functions to preserve mode type and representational content, without adding warrant or new subject matter to current psychological transactions. (See note 63.) Fulfilling this function by purely preservative memory is constitutive to *any* representational psychology—propositional or not, meta-representational or not. For example, in propositional inference the mode and content of earlier steps must be preserved for later use in inference. Or in a use of visual perception in action, purely preservative memory retains the mode and representational content of the visual state. Although one is entitled to rely on purely preservative memory, and although such reliance is necessary for the success of the inference or perception-based action, such reliance is not part of the justification or a practical warrant for the action. Warranted reliance on purely preservative memory is a necessary enabling condition for the conclusion's, or action's, being warranted. My account of the epistemic warrants for Self-Understanding and Self-Knowledge in non-*cogito* cases hinges on reflecting on preservational capacities, central to any representational psychology, and their role in meta-representational understanding—especially betokening understanding.

There are importantly different types of Self-Understanding and Self-Knowledge. Preservation, though not purely preservative memory, figures in all of them. Preservation is necessary to Self-Understanding from the inside, and to the role of Self-Understanding in providing buck-stopping status to psychological elements in critical reason and morality. It is part of the constitutive tissue of self-hood. So understanding the forms that preservation takes between lower-level and meta-psychological elements lies at the heart of my project.

---

[98] I write 'preserves sensitivity' because the lower-level purely preservative memory does not preserve any representation of a belief as such. Lower-level aspects of a psychology are not meta-psychological. They nevertheless track differences in the modes of past steps in inferences. Beliefs function differently in the psychology from suppositions. Purely preservative memory tracks sensitivities to different modes, and in that sense preserves the mode types.

In all cases, preservation between the lower-level understanding and the meta-representational understanding *of the representational content* works in the same way. In propositional cases, the lower-level thought content is always preserved in the meta-representational thought by being *used*—thought—*in* the canonical name for the content. I will concentrate on various types of preservation of *mode type* and *mode instance*.

Preservation between lower-level thought and meta-representation of it occurs trivially in pure *cogito* cases. No memory is involved. In pure—formally self-verifying—*cogito* cases (I [hereby] entertain the thought that writing requires concentration), the mode—entertaining—of the lower-level thought (writing requires concentration) is inevitably preserved in any thinking of the meta-representational thought, no matter what its mode is. Entertaining the lower-level thought content is guaranteed in the meta-representational thought. Even the instance of entertaining is referred to reflexively in the thinking of *cogito*. Pure *cogito* cases are infallible. If one thinks a thought of that form, one thinks a truth.

In impure *cogito* cases (I [hereby] judge that writing requires concentration), the indicated lower-level mode (judgment) is not entailed to be the mode of the content cited by the underlining. In such a *cogito* thought, one could judge that one is judging the lower-level content, and be mistaken about one's making a judgment. The two levels of judgment could come apart. One could judge that one is judging that writing requires concentration, although one does not thereby *judge* that writing requires concentration. The latter content (that writing requires concentration) might pass noncommittally through one's mind. Such cases are pathological misuses of the *cogito* form of thought. Normally, the meta-representational judgment coordinates with a performative judgment of the lower-level content. Normally, the lower-level judgment is a constituent of the meta-representational act of judgment. Then the lower-level mode and its instance are preserved in the meta-representational judgment not by entailment, but by coordinated, performative, reflexive agency.[90]

Cogito judgments concern present occurrences. Preserving mode type and mode instance is trivial in pure cases, and almost trivial in impure cases. It does not depend on a causal connection between the understood psychological element and the understanding. The connection is contemporaneous. Preservation of mode in judgments that realize Self-Understanding of one's present, continuing, *standing* attitudinal *states* (belief, intention) and certain among one's past psychological states and episodes is more interesting.

---

[90] For more on *cogito* cases, see my "Individualism and Self-Knowledge"; and "Mental Agency and Authoritative Self-Knowledge: Reply to Kobes," in Martin Hahn and Bjørn Ramberg, eds., *Reflections and Replies: Essays on the Philosophy of Tyler Burge* (Cambridge; MIT, 2003), pp. 417–34. In the latter article, I discuss impure performative *cogito* judgments involving modes like intention as well as modes like judgment. See note 74.

How does preservation work in Self-Understanding of standing belief? Memory often figures in a present-tense judgment like I believe that Mozart loved Haydn. Purely preservative memory retains a standing belief—standing from before the judgment. The judgment takes the belief not to have just formed. Purely preservative memory preserves content and mode of the belief for later use. In a meta-representational psychology, the lower-level, purely preservative memory is supplemented by conceptualized representation of both the preserved representational content and the preserved lower-level mode—here, belief. In conceptualizing these matters and *representing* the belief as the individual's own, the individual indicates not only the mode type *belief*; the individual also thinks *de re* of the mode instance. The individual represents his or her own instance of that type of belief in the judgment. When combined with canonical understanding of representational content, such powers yield betokening understanding of instances of standing propositional states.

The memory used in this sort of recognition is meta-psychological autobiographical memory. But it is not episodic memory. The memory is purely preservative memory, supplemented with a meta-representational capacity (a) to conceptualize the tracked preserved mode—type and instance—and representational content of the standing state, and (b) to recognize that it is a state that has been standing from the past.

A similar account applies to Self-Understanding of some past psychological events, especially relatively recent ones. Meta-psychological autobiographical episodic memory again relies on purely preservative memory, supplementing it by conceptualizing the representational content and mode (type and instance) of the past psychological event.[100]

Some Self-Understanding and Self-Knowledge are forward looking. One can be entitled to believe, fallibly but nonempirically, that one will retain a present intention or belief some moments, or longer, into the future. One can anticipate that a specific standing state type and state instance will remain standing into at least the near future, if there is no reason to doubt. Forward tracks of preservation are as basic to representational psychologies as backward, causally sustained, memory tracks. When utilized by meta-representational powers, including canonical representation of representational content, they figure in *de re* understanding of the mode type and mode instance of future representational states.

---

[100] Preservation of mode in memory need not maintain commitment involved in the mode. One might have a meta-psychological episodic memory of one's having judged that so and so, even as one currently doubts so and so. Past attitude instances are preserved, inasmuch as betokening understanding relies on lower-level preservational memory to track that commitment.

Use of these types of preservation—of a present event, of a standing state or a past episode through purely preservative memory, and of a future anticipated state or event—is constitutive to being a self. An individual that lacked any one of these preservational capacities would lack powers integral to applying critical reason in inference or action. I made an analogous point in Lecture II regarding the role of meta-psychological autobiographical episodic memory in critical inference, decision making, and dialectic.

VI

I turn now to *epistemic warrant* for betokening understanding in Self-Understanding. Epistemic norms are grounded in competencies to realize representational functions. Realizing representational functions involves forming and preserving veridical psychological states and occurrences. A propositional attitude is epistemically warranted only if it is the product of a competence that realizes its representational function reliably and well, given the individual's perspective. Powers to produce and preserve veridical representation—representation of mode types and mode instances, and of representational contents—must, in their natural state, be reliably veridical if they are to be constitutive aspects of selves and help ground critical reason. So reliance on these powers by selves—critical reasoners—is epistemically warranted. In particular, in purporting to single out *de re* one's propositional attitude instances, one is default warranted in betokening understanding, if one bases judgments about the instances on the relevant types of preservation.

The judgments based on the preservational routes just discussed are immune to brute error. The meta-representational understanding relies on lower-level preservational capacities in specifying representational content, mode type, and mode instance. The warrants that yield immunity to brute error are all based on this reliance of meta-representational understanding on the lower-level competence understanding, perhaps supplemented by purely preservative memory or anticipation. I will apply these epistemic points to each of the types of Self-Understanding that I have discussed.

Pure *cogito* thinkings are warranted through self-evidence. Individuals are warranted by understanding their thinkings. Such warrants yield immunity to *any* error. Warrant for impure *cogito* judgments lies in *de re* understanding. Understanding coordinates the lower-level performative act with the higher-level *cogito* judgment. Impure *cogito* judgments are not immune to error. Coordination failures that yield error are pathological and undermine warrant. If one judges I am hereby judging that Mozart loved Haydn and fails to judge that Mozart loved Haydn, one's self-understanding is pathological; warrant lapses. One misuses such judgments if one does not engage in the act that one judges oneself to be engaging in. Such misuses undermine warrant. Impure *cogito* cases are immune to brute error.

Errors in judgments, based on Self-Understanding, about standing propositional attitudes, past or present, derive from malfunction of purely preservative memory or from relying on other capacities besides purely preservative memory. Then the judgment either is unwarranted or is backed by a different warrant from the one that backs Self-Understanding.

Suppose that one applies the concept intend to finish the project or the concept believe that most domesticated dogs are friendly in a purported betokening understanding. Suppose that one lacks the attributed psychological state. The first-level purely preservative memory could have malfunctioned. Or it could have been derailed by competing psychological needs, or other sources of self-deception. Malfunction or relying—however unconsciously—on factors that are irrelevant to supporting a belief undermine the relevant warrant (entitlement). Or one could form a warranted but mistaken judgment with the same content, inferring it from behavioral evidence. Then one's warrant would be different. In all relevant cases of error, either one's default entitlement for the judgment about one's standing propositional attitude lapses, or the judgment does not rest on the *relevant* entitlement. So there is no brute error in judgments warranted in the relevant way. The key route is a conceptualization of mode and content that is preserved by purely preservative memory. The capacity to track one's standing states by way of their mode and content, and to represent them as having been held (not presently formed), is a basic preservational power in a critical reasoner. Minimal exercise of this power yields immunity to brute error.

A similar line applies, I think, to some judgments about past psychological events—judgments based on meta-psychological autobiographical episodic memory that relies on lower-level purely preservative memory. Suppose that one judges I was just then thinking that writing requires concentration. The judgment rests on meta-psychological autobiographical episodic memory. The memory conceptualizes a mode and content preserved by the basic trans-temporal psychological connector—purely preservative memory.

I claim that relying on betokening understanding that utilizes purely preservative memory in Self-Understanding of both some standing propositional attitude states and some past propositional attitude events is immune to brute error. I want to elaborate this claim and defend it. First, again, I discuss the warrant for relying on purely preservative memory in these types of Self-Understanding.

Purely preservative memory functions just to retain the mode and representational content of psychological states. It does not represent anything. It only connects. It preserves representation from other states. Relying on it is warranted nonempirically, though, of course, it can preserve empirical content. It is a constitutive precondition of any inference. Its being naturally reliable is a condition on having a functioning representational psychology.

Purely preservative memory is fallible. It can lapse momentarily, or become unreliable, through age or disease. Short-term forms do not operate after their decay times. Longer-term forms, which, as far as is known, lack built-in time limits, not only fail on occasions but deteriorate in reliability. Still, whenever purely preservative memory fails to retain mode or content, the failure is a malfunction. Errors are not attributable to a representationally well-functioning preservative competence. Any malfunction undermines warrant.

Of course, error can derive from other psychological factors' overriding purely preservative memory, or being used instead of it. Then errors do not derive from purely preservative memory. Since the individual relies on other factors, warranted or not, any such errors do not occur despite warranted reliance on purely preservative memory.

How and whether purely preservative memory interacts with other psychological factors, in general or on any given occasion, is an empirical matter. But when it does not malfunction and is not overridden by other psychological factors, it preserves mode and content of past representational states. Where there are errors that are attributable to *its* operations, the errors are not brute errors. They are errors of malfunction that undermine warrant. That is why reliance on purely preservative memory is immune to brute error.

Are other types of memory, besides purely preservative memory, subject to brute error? In a sense, perceptual memory inherits perception's—and perceptual belief's—vulnerability to brute error. A given memory could be brute mistaken because it derives from a perception that is brute mistaken. However, here, the vulnerability is the vulnerability of *perception* to brute error. Errors specifically attributable to memory can seem to result from memory's not making the best cognitive contribution that it can. And it may seem that when perceptual memory undergoes such failure, warrant for relying on it lapses, because memory would on such occasions not operate well, even if it is generally reliable. Such failures would differentiate memory from perception. Perception can operate optimally given its proximal input; but because of abnormal distal conditions, it can still go wrong. Cases in which one is warranted in relying on it are cases in which it operates well, given the information that it has. If inaccuracy is correctly attributable to the operation of the perceptual system in a given case, warrant for relying on perception lapses. It can seem that where error isattributable specifically to perceptual memory, warrant lapses. So it can seem unclear how brute error in *any* memory is possible.

This account is too simple. It is known that some memory—especially propositional, longer-term episodic memory—involves constructing narratives and carrying out implicit (unconscious) quasi-inductions about the past. These constructions are partly based on other things that we believe. They are fairly reliable in providing important general attributes of remembered episodes. But with respect to matters in the relatively distant past, memory is often strikingly unreliable in filling in specific details.[101]

One can make sense of brute error attributable to such memory. To be warranted, the memory type must be reliable with respect to the level of detail remembered in the case in which the memory is employed. One would not be entitled to rely on purported memory of certain types of details that are purportedly remembered in the distant past, if memory in such cases is not reliable. But in cases in which memory is reliable in constructing the general attributes of an episode, it could fall into brute error despite following warranted operations from warranted starting points.

I make three points regarding this conjecture about brute error in episodic memory. First, to count as memory of an episode, a constructionist memory must use not only general background beliefs, but nonconstructed, non-quasi-inductively produced representations of the episode. (The nonconstructed memories are the singular bases for the quasi-induction.) I doubt that these singular starting points for the memory construction could be subject to brute error. Either they are well formed and accurate, or they are inaccurate by virtue of a nonconstructed retention failure by the generally reliable competence. Such retention failures are just the sort of failures in representational function that undermine warrant.

---

[101] See Daniel Schachter, *Searching for Memory The Brain, the Mind, and the Past* (New York: Basic, 1996). Schachter notes that John Dean's, presumably honest, detailed accounts of particular episodes were shown by the Nixon tapes to be massively inaccurate regarding details of the episodes, but accurate on their most important general features.

Second, if there were no memories—whether or not these are singular bases for constructionist memories—of one's near-term propositional states and events that were immune to brute error, one's reflexive reviews of such states and events would never give the reviewed states and events buck-stopping status in critical reasoning and moral reasoning—by the arguments I gave in section IV of Lecture II and section I of this lecture. We have good reflective reason to believe that reflective review does give reviewed states buck-stopping status in uses of inference, practical decision making, and dialectic.

Third, the idea that near-term episodic memories always depend on matters beyond the well-functioning of one's memory—and are always subject to brute error—is implausible on its face. The idea that I was just thinking then [five seconds ago] that writing requires concentration or I am carrying out my decision [just made] to help him is always subject to error that could result from memory that functions well in the given circumstances is implausible on its face. Assume that the thought is based on conscious memory of the event, and not some collateral conscious thoughts about the past. To be subject to brute error the memory-based thought must be warranted. So the episodic memory must be generally reliable in that type of case. Error must not derive from contextual failure in the well-functioning operation of memory. For such failure would undermine warrant. Holding that memory of near-term past thinkings is always subject to errors that cast no aspersions on the representational well-functioning of memory is extremely implausible.

I think that our understanding of diachronic reasoning in inference, practical decision making, and dialectic indicates that in such reasoning episodic memory can utilize purely preservative memory—a type of what is called 'semantic memory' in the psychological literature. Such memory is the nonconstructionist, anaphoric type of retention required in any diachronic representational operation—including inductive inference and narrative. Our understanding of the epistemic norms governing memory-based meta-psychological thoughts about one's very recent thinkings indicates that errors in such thoughts are attributable to representational malfunctions of memory. Such malfunctions, on given occasions, undermine warrant. Since I think that a natural competence to retain thought content and occurrent thought events over at least short times is constitutive to critical reason in temporal beings, I think that in us, and in any critical reasoner, there are cases of near-term meta-psychological autobiographical episodic memory in which brute error is not possible. I think that any individual that is subject to brute error in all memories of near-term thinkings would lack the cognitive equipment that is constitutive to the use of critical reason in inference, practical decision making, and dialectical thinking.[102]

I have been discussing warrants for relying on purely preservative and episodic memory in Self-Understanding of standing propositional states and past propositional events. Those warrants are aspects of the warrant for betokening understanding in these cases. I want now to focus more specifically on the warrant for betokening understanding in these types of Self-Understanding. In these cases, betokening understanding relies on purely preservative memory in Self-Understanding to retain mode and content of standing or past psychological elements.[103] Betokening understanding supplements this reliance with use of meta-representational competencies to *represent* the preserved representational content in a canonical way and to *represent* the preserved mode instance.

Purely preservative memory preserves mode type and representational content, but *represents* neither as a subject matter. *A fortiori* it does not represent the events or state instances whose mode and content it preserves. Still, it always preserves the mode type and representational content *of a particular antecedent state instance or event.* In Self-Understanding, betokening understanding capitalizes on this fact to represent not only the mode type and representational content, but also the psychological state instance or event—which is always preserved, "there" to be referred to, when purely preservative memory functions well. Betokening understanding *uses* the lower-level purely preservative memory while *representing* what it preserves. This meta-representation is *de re* and is immune to brute error. So, when betokening understanding relies on just these competencies, either it is warranted and veridical, or it goes wrong through malfunction and warrant is undermined. Thus the basic warrant for relying on betokening understanding is immune to brute error.

---

[102] If it is possible that there be beings that carry out critical reason in time periods shorter than those allowed for in current views about long-term memory—of which episodic memory is a subtype—then I think that a competence for an analog of meta-psychological autobiographical episodic memory would have to occur in those beings. The analog would differ only in being short-term memory. It would be memory of reasoning episodes in the very *very* recent—technically, short-term—past. Those analogs would have to be immune to brute error. Any being with representational competencies must, I think, have at least short-term memory. And beings with critical reason must be able to utilize autobiographical meta-psychological memory that retains not only content in reasoning, but retains *de re* memory of episodes of reasoning, individuated by their mode and representational content. Whether the memory is technically long term is, I think, not a constitutive matter. But in all actual critical reasoners that we know of, the relevant memory is surely sometimes long term—a few seconds or longer.

[103] In Self-Understanding of past psychological events, purely preservative memory joins, as we have seen, with episodic memory. In what follows, I focus on the relation between betokening understanding and purely preservative memory. It should be understood that what I say is meant to carry over to the cases of Self-Understanding in which episodic memory is involved as well.

What are the epistemic bases for my claim that betokening understanding in Self-Understanding of *some* propositional attitude states and events is immune to brute error?

I noted that we know nonempirically, by understanding their reflexive structure, that pure and impure *cogito* judgments involve no inference. Pure *cogito* cases are infallible. By reflecting on the form and content of the judgments, one can recognize that in impure *cogito* cases only a pathological failure to hold meta-representational and lower-level attitude modes together in a judgment could yield error, if one thinks the thoughts in a reflexive, "hereby" way. Errors derive from odd misuses of a form that one can use successfully. Although these are easy cases, they provide a model for understanding a wider range of cases.

What is the epistemic basis for my meta-meta-claim that betokening understanding in Self-Understanding of *some* past propositional events, and *some* standing attitude instances, as standing, is immune to brute error? The betokening understanding consists in a meta-representational conceptualization of the products of purely preservative memory.

The point that relying on *purely preservative memory* is immune to brute error seems secure for reasons already given. The point depends on reflecting on the function of such memory in any representational psychology. A reliable capacity to preserve mode and content is constitutive to having a system that could count as a representational psychology. Purely preservative memory is certainly not a matter of induction. Like deduction, induction *presupposes* a prior noninferential capacity for purely preservative memory. The retention does not involve representation of what is retained. Error in retention must consist in representational malfunction of the competence, not brute mismatch with a subject matter.

What of the epistemic status of meta-psychological conceptualizations of representational content and mode (type and instance) of the products of purely preservative memory? Canonical meta-representational conceptualizations of *representational content* seem to be exercises of a competence understanding of the content together with competence understanding of the that-clause-like canonical names. Those are competencies that we clearly have. Reflection on them shows at least as clearly as reflection on *cogito* competencies does that they are nonempirical and immune to brute error. The key issue concerns the status of my claim regarding conceptualization of *mode—type* and *instance*—of a state with a specific representational content in meta-psychological memory. How can we know that such conceptualization is immune to brute error?

In Lecture II, section IV, I discussed three ways in which meta-psychological autobiographical episodic memory from the inside figures constitutively in critical reasoning: in inference, in practical decision making, and in dialectic. In such exercises of critical reason, we can recognize conceptualization of a mode—type and instance—of a propositional attitude that is preserved for a rational project (inference, decision, dialectic).

The following principle is a norm of critical reasoning in these enterprises:

> (CR) If, in critical reasoning, one correctly and with warrant judges that a lower-level state is (or is not) reasonable, then it rationally follows directly that one has reason to sustain (or change) the lower-level state.[104]

(CR) is like an inference rule, except that it explicitly concerns transmission of reasons, not preservation of truth. (CR) implies that in the relevant uses of critical reasoning, one's Self-Understanding of the lower-level state instance is immune to brute error. For if an application were subject to brute error, a reason that is attributed, correctly and with warrant, at the meta-psychological level of critical reasoning would not transmit immediately to the lower-level state. The transmission would *always* be subject to the proviso that one's completely warranted meta-level evaluation and self-attribution is not brute mistaken. If brute error were always possible in one's self-understanding, the lower-level attitude would not have buck-stopping status by virtue of being reviewed or reviewable. In fact, *no* attitude, except in *cogito* cases, would have buck-stopping status by virtue of being reviewed or reviewable. In critical reason, the fundamental, buck-stopping states include both the meta-level review and the critically reviewed lower-level states. The buck-stopping status of the lower-level states in critical reasoning partly depends on the potential for meta-level review. They are not separable. If (CR) were not true and applicable in the three cases, critical reason would never occur outside the immediate present of *cogito* cases. Thus I think that the view that understanding our standing propositional attitudes, as having stood, and understanding of some—at least near-term—past propositional occurrences is *always* subject to brute error amounts to a skepticism about our being diachronically extended critical reasoners.

---

[104] See my "Our Entitlement to Self-Knowledge," p 109ff.

I think that we can know that (CR) governs our reasoning in the way that we know that our reasoning is governed by specific rules of deductive inference—or by specific principles governing simple deductive reason-support relations. Knowing the principle is based on reflecting on the practice and norms of critical reasoning.

I think that we can know, by reflection, (CR)'s application to some actual particular cases of critical reasoning. That is, I think that in some cases, we can know by reflection that, from a warranted meta-level judgment that a given attitude is reasonable or unreasonable, we immediately have a reason to maintain or change the lower-level attitude. We have a reliable and warranted capacity to recognize the form of transmission of reasons as they occur in some diachronic cases. This capacity is analogous to recognizing in a deductive inference, on a particular occasion, that if a premise is warranted, it provides a reason for the conclusion.

Our ability to recognize immediate reason transmission in given cases is fallible and defeasible. But it seems to me that, as in the case of knowing (CR) itself, the positive warrant for knowing some of its applications is nonempirical. The warrant is for applications of a meta-meta-representational capacity, a capacity to understand our tracking representational content and mode (type and instance) of our diachronic conscious reasoning.

It is true that the unconscious mechanisms that underlie memory in these cases are open only to empirical investigation. But the conscious states are not identical with unconscious counterparts, even those that share *representata*. For the modes of presentation and hence identities of conscious states are different. Consciousness is an aspect of the mode of presentation, hence identity, of the states and processes that I am centering on. I think that the idea that we can know nonempirically the form of reason transmission in particular cases is no less plausible than the idea that we can know nonempirically the form of deductive inferences in particular cases, or that we can know of one type of thought that it is, if warranted, a reason for another. Such knowledge is defeasible.[105] But it is apriori. The idea that we have it seems to me very powerfully plausible. Such knowledge of norms places a prima facie constraint on empirical psychological investigations.

---

[105] The claim that we have such apriori knowledge is defeasible. Most apriori warrants for Self-Knowledge are themselves defeasible. Hume was skeptical of reasons, as opposed to natural associations He tended to assimilate all reasoning to induction, which he construed as blind habitual association. I join the later history of philosophy in rejecting these doubts. I think that some meta-knowledge that we engage in deductive inference is probably not defeasible.

Not all memory in self-knowledge about past psychological states is immune to brute error. Most memories with warrants immune to brute error are nearer term, even if they are technically long-term memories. As time spans lengthen, use of induction about the past and memory that makes essential use of cobbling together different experiential memories become more prevalent. Longer-term memories often rely on factors other than purely preservative memory to access past psychological events—factors subject to brute error. My thesis is that we, as selves and critical reasoners, do sometimes rely on meta-psychological autobiographical episodic memory, supported by purely preservative memory, in calling up past thoughts. I leave open in what cases we do so.

In sum, a betokening understanding that connects *de re* applications to preserved mode instances, and to their representational contents, relies on a lower-level capacity, purely preservative memory, that is constitutive to reason. The betokening understanding itself is a constitutive connector in a psychology capable of *critical* reason.[106] An individual is entitled to such betokening understanding, and to judgments based on it, because (a) he or she is entitled to rely on the lower-level capacity, and (b) he or she is entitled to *de re* meta-conceptualization of the mode and content of attitude instances used in inference, decision making, and dialectic in critical reasoning. Moreover, we have a meta-meta-representational ability to recognize applications of the capacity cited in (b). Judgments that rely on this betokening understanding are immune to brute error, because any error in such judgments derives from failure of purely preservative memory, failure of a *de re* power of understanding constitutive to critical reason, or failure to base the judgment on those capacities.[107]

Analogous points apply to anticipatory betokening understanding. Such understanding of future psychological instances is meta-representational conceptualization that piggybacks on lower-level preservational capacities for control and continuity in states like intentions—for example, in decision making. The default entitlements attach to an understanding that is a basic connector in critical reason. They yield immunity to brute error.

Let me survey whence we have come. I have focused on betokening understanding of propositional states and occurrences. Such understanding is integral to an apperceptive rationally accessible point of view, hence to being a self. Apperception combines betokening understanding with a self's self-attribution of the understood psychological elements. It comprises the multi-tier perspective of critical reason that marks selves. Betokening understanding is also essential to beings that are subject to moral norms. It is a meta-representational capacity that, except in *cogito*-like cases, uses lower-level trans-temporal preservational powers to guide *de re* representation of propositional state instances and propositional occurrences. Natural reliability of the lower-level preservational capacities is constitutive to any representational psychology, hence to reason. Natural reliability of the meta-representational betokening understanding that uses those lower-level capacities preserves Self-Understood state instances and events in a way that is constitutive to selves and to psychologies capable of critical reason and morality. Such preservation is basic to understanding one's states from the inside and to the buck-stopping status of Self-Understandable states in critical reason and morality. Epistemic warrants attach to betokening understanding because it is reliable. Successful representation, on particular occasions, depends only on its good use. It does not depend on matching a subject matter outside the psychology, or on information beyond the minimum necessary to exercise the understanding. Being warranted follows from using a well-functioning, reliable, epistemically viable competence, relative to information available to it. (See (E) above.) Such uses are certainly epistemically viable and, within their appropriate domain, reliable. And the minimum information needed to use betokening understanding suffices to provide veridical memory. So when the understanding does not malfunction and is not misused, it is warranted. Error derives only from malfunction, from misuse, or from relying on other capacities. So betokening understanding is immune to brute error, relative to this warrant.

---

[106] See discussion of meta-psychological autobiographical episodic memory in Lecture II.

[107] I reemphasize that this failure could involve relying on inference or on considerations that are relevant to supplementing or overriding the basic entitlement.

Fulfilling epistemic norms governing Self-Knowledge thus entails exercising powers constitutive to *critical reasoning*. They are not norms that are merely necessary to critical reasoning, but constitutively independent of it. Fulfilling them requires using preservational powers that are partly constitutive of critical reason. For betokening understanding of attitudes constitutively uses meta-representation guided by preservational powers that are constitutive to any reasoning.[108]

---

[108] These points and the next text paragraphs address an interesting objection by Peacocke, "Our Entitlement to Self-Knowledge: Entitlement, Self-Knowledge, and Conceptual Re-Deployment," *Proceedings of the Aristotelian Society*, XCVI (1996): 117–58. Peacocke objected to my account in "Our Entitlement to Self-Knowledge." He held that

Conversely, such understanding is constitutive to the multi-tier, apperceptive point of view that grounds applicability of norms of critical reason.[109]

VII

I now discuss a further aspect of betokening understanding. Critical reasoners must use understanding of lower-level reason-support relations to guide betokening understanding of instances of propositional attitudes. Betokening understanding must be guided not only by competence with lower-level preservation relations, but also by conceptualizations of reason-support relations among the target attitude and other lower-level attitudes. Such relations place the target attitude instance in a network of reason relations. (See note 108.) Any

---

I had successfully shown only that critical reason requires Self-Knowledge, not that critical reason figures constitutively in the Self-Knowledge. He thought that Self-Knowledge is constitutively independent of norms of critical reason, and that mastery of the concept of belief and other attitudes, mastery of concepts of representational contents, and sensitivity to first-order norms of reason "generate" an explanation of Self-Knowledge. My 1996 paper did not focus on this issue; and Peacocke's objection had some resonance, although I believe that my account of the structure of critical reason even then suggested why it was plausible to hold that Self-Knowledge in critical reasoners uses essential preservative routes that are distinctive to critical reasoning. I have maintained that rational capacities that figure in critical reasoning are partly constitutive of Self-Knowledge.

Of course, I do not hold that Self-Knowledge is *based on* critical reasoning. Except in *cogito* cases, it is not based on reasons at all. My view is rather, first, that preservative routes essential to any reasoning, hence critical reasoning, are essential to Self-Knowledge. So Self-Knowledge cannot be constitutively independent of norms of critical reason. Perhaps this point is implicit in Peacocke's own appeal to sensitivity to first-order norms of reason. Second, I think that competence to conceive reasons as such is partly constitutive of mastering concepts of belief and other attitudes. See the discussion, in section VII below, of using reason-support relations as a supplementary factor in betokening understanding. Both points bring out ways in which horizontal and vertical reason-preservation and reason-transmission routes that are constitutive of critical reasoning also figure constitutively in Self-Knowledge. For further discussion, which however does not advance as far as the present account, see my "A Century of Deflation and a Moment about Self-Knowledge," Presidential Address to the Pacific APA, *Proceedings and Addresses of the American Philosophical Association*, LXXIII, 2 (November 1999): 25–46.

The account here also answers objections to my "Individualism and Self-Knowledge," by Sven Bernecker, "Externalism and the Attitudinal Component of Self-Knowledge," *Noûs*, XXX, 2 (June 1996): 262–75. Bernecker noted that "Individualism and Self-Knowledge" did not explain warrant in knowing the modes of our attitudes (belief rather than supposition, intention rather than hope). As I have emphasized elsewhere, "Individualism and Self-Knowledge" was not intended as an epistemic account of Self-Knowledge, much less self-knowledge.

[109] I presented this account in the oral form of these lectures in December 2007 at Columbia University. The general outlines of this account of one's warrant for Self-Knowledge are adumbrated independently by Frank Olav Barel in his dissertation *Self-Knowledge and Anti-Individualism: A Defence of Neo-Rationalistic Compatibilism* (University of Oslo, 2010).

understanding of attitude instances that could not conceptualize reason-support relations as such could not ground the applicability of norms of morality and critical reason. In betokening an attitude instance, Self-Understanding essentially uses *de re* understanding of specific reason-support relations, practical or theoretical, among that attitude and others. Understanding an attitude's place in a network of reasons, for or against, bears on epistemic entitlement to judgments based on betokening understanding of that attitude. Since the identity of propositional attitudes constitutively depends partly on reason relations to other attitudes, betokening informed by understanding such connections is a better route to veridicality than any would-be betokening not so informed. So this identification through reason support in betokening understanding contributes to warrant for Self-Knowledge.

As with betokening understanding, lower-level competence understanding is used *in* the meta-representational understanding of reason-support relations as such (*R* is a reason for *A*). Understanding reason-support relations requires non-meta-representational competence to *use* one content as reason for another, and some disposition to be moved by, to *implement*, relations of threat or support. To understand a reason as such, for or against an attitude, one must have some lower-level disposition to shift or maintain attitudes in accord with the reason. Recognizing a reason as such essentially relies on the lower-level competence to use the reason to support the conclusion. This competence figures in the conceptual mode of presentation of the reason-support relation between a given betokened attitude and others.[110]

Take the concept, is a reason for believing that someone loved Haydn. Suppose it true of an instance of a belief that Mozart loved Haydn. The betokening understanding of that belief might be partly guided by an understanding of its supporting a belief that someone loved Haydn. Understanding the instance's supporting and being supported by other possible attitudes helps guide the betokening understanding. Concepts of reason support are meta-representational. In using them, however, a thinker must rely on a lower-level competence understanding of the *represented* propositional attitude types—belief that Mozart loved Haydn and belief that someone loved Haydn—including a lower-level competence to reason from one to the other. One relies on lower-level understanding of the represented contents as part of thinking about those contents, and about reason-support relations among them. The key point is that the guiding conceptualizations of reason-support relations employ lower-level competence understanding of the targeted contents, modes, and reason-support relations.

---

[110] Cf. "Reason and the First Person."

This point supplements, I think constitutively, the account already given of immunity to brute error yielded by the default warrant for betokening understanding and judgments based on it. Meta-representational understanding of reason-support relations must use lower-level competence understanding. The lower-level competence understanding is warranted through correctly representing the relations of reason support that set norms for epistemic warrant. The meta-representational understanding is warranted by relying on correct lower-level competence understanding of reason-support relations. So a warrant for meta-representational understanding of reason-support relations yields immunity to brute error.

To summarize the idea crudely, meta-representational understanding of reason-support relations depends for its warrant on sucking up warrant, present in the lower-level competence, for engaging in lower-level reasoning. To be warranted, the meta-representational understanding must involve correct understanding, and involve a disposition to be moved by, the lower-level reason-support relations. The meta-representational understanding of reason-support relations helps guide betokening understanding of attitude instances by locating them in a local network of practical and theoretical reasons. Meta-representational understanding of such relations improves one's route to veridical betokening understanding. Meta-representational understanding of reason-support relations contributes warrant to betokening understanding by helping it follow this route, only if it is correct understanding. It must be understanding of relations of genuine reason support. So both the warrant for meta-representational understanding of reason-support relations and the contribution of warrant for such understanding to the warrant for betokening understanding yield immunity to brute error. Clearly, this aspect of the warrant for betokening understanding is constitutive to critical reason—not merely a component of critical reasoning whose warrant comes from outside critical reason itself. (See note 108.)

VIII

I turn to betokening understanding in Self-Understanding and Self-Knowledge of two types of *nonpropositional* psychological states and occurrences. One type comprises nonpropositional *representational* empirical states—such as perceptions and perceptually derived images. The other comprises conscious sensory states, such as certain aspects of pains, that are not representational in my sense. I will have to deal with these cases briefly.

The account of betokening understanding of perceptions, perceptual anticipations, perceptual memories, and fictional images broadly follows the account of betokening understanding of propositional states. Capacities to preserve mode and content of perceptions are constitutive to any empirical representational psychology. In a propositional-attitude psychology, such preservational connections obtain between perception and first-order perceptual belief. If a psychology has meta-representational capacities, some of them conceptualize lower-level contents *in* specifying those contents. Just as object-level perceptual concepts use perceptual contents in their modes of presenting environmental entities, so meta-concepts that represent perceptual contents—especially perceptual attributives—employ the object-level perceptual contents in *their* modes of representing those contents. Thinking about perceptions also employs lower-level mechanisms for preserving perceptual mode in conceiving of that mode, type and instance. Meta-representation that uses lower-level capacities is constitutive to the apperceptive core self.

An individual is default entitled to beliefs based on reliably veridical betokening understanding. Reliability is necessary for there to *be* the relevant point of view. Errors derive only from malfunction or misuse of the preservational operations—undermining entitlement—or from failure to base belief on those operations. A relevantly based judgment cannot be mistaken about a current perception, or immediately past perception, or a perceptual anticipation, while all is well with the cognitive apparatus. Brute error—warranted, representationally well-functioning error—is not possible relative to the relevant warrant.

Betokening understanding of *nonrepresentational* conscious sensations, or aspects of them, differs importantly from the other types of betokening understanding. No lower-level capacities preserve representational content. But the basic shape of the account of the warrant for Self-Understanding of conscious sensations is otherwise similar.

Phenomenal consciousness of sensations is constitutive to being a self. Rational access to such consciousness is also constitutive.[111] Here betokening understanding forms concepts of sensations. The concepts constitutively use the sensations, or memories of their types, in representing them. So a lower-level element is again used in conceptual representation of it. A capacity to use such concepts *de re* to pick out instances is not constitutive to reasoning. But it is constitutive to being a self. A self must be able to think about and act on some conscious affect, such as its pains or pleasures. Such a capacity is basic to owning and controlling bedrock aspects of conscious life—the core of consciousness in a point of view. Being subject to moral norms also seems to hinge on this capacity.

---

[111] Higher animals and very young children that lack selves can surely form beliefs about sensations like pain. Such beliefs are as primitive as ordinary perceptual beliefs about the environment. Of course, such beings cannot *self*-attribute having those sensations; for they are not selves. But their beliefs make the sensations ego-relevant through whatever ego-concept they have that precedes a full-fledged self-concept. Traditional theology might allow that God is a self and lacks sensations. I confine my thesis to finite selves that interact with the physical world.

A reliable natural capacity to form true beliefs in normal circumstances—those in which the beliefs' representational natures were formed—yields epistemic entitlement to those beliefs. Normally, not all is cognitively well with a person who mistakenly believes he or she is in sharp pain, or that a sharp pain is a tickle. Error can, of course, arise from anticipation, inference, or ideology. But any error that relies on the most direct conscious way of thinking about a sensation is a dislocation of self—a pathology that undermines warrant.[112] Beliefs about one's conscious sensations are immune to brute error, *relative to the entitlement*.[113]

This completes my sketch of applications of betokening understanding, and judgments based on it, to nonpropositional cases.

IX

Predications in Self-Understanding are *self-attributions of* betokening understanding. Self-attribution is attribution of the psychological state to oneself via self-conception.

What is the warrant for the transition from warranted *de re* betokening understanding to self-attribution? Wherein is one entitled, in Self-Understanding, to attribute *to oneself*—referred to as oneself, with a self-concept—an instance of an understood state?

---

[112] There are borderline cases. Perhaps some pains are so close to borderlines with other types of sensation that, even in ordinary circumstances, one could make a brute error about whether one is having one. (Presumably, one could still get right that it is a sensation of a more generic type.) Perhaps one could even be *unreliable* in forming judgments about certain types of sensations that are easily mis-categorized, at least given only short periods of time. Such cases do not, I think, count against the view that many judgments about conscious sensations are immune to brute error relative to certain warrants associated with relevant judgment-forming routes.

Again, I think that whether a warrant yields immunity to brute error is not apriori determinable from the content of the belief. *A fortiori*, I think that immunity to all error is not intrinsic to self-attributions with sensation concepts. I think that because of cognitive malfunction or stoic ideology, an individual can judge even a very intense pain not to be a pain.

[113] I think that *some* judgments about *past* sensations, and perceptions, are immune to brute error relative to certain warrants. The relevant warrants attach to judgments that use purely preservative memory of singular applications of concepts of those sensations or perceptions. Here the account parallels that of betokening understanding of past propositional thought events.

The transition is idealized. There is no step in time. Betokening understanding and self-attribution are aspects of the predication. The relation between the aspects is, however, usefully considered as a transition step, although there is no literal inference. It is *not* an inference from That is a belief that *p* to I believe that *p* (or that belief is mine).

The epistemic norm for the transition step, governing warrant for basic apperception is as follows:

> (WBA) If one is entitled in the relevant way to betokening understanding of an instance of a psychological element, and if the self-attribution relies only on the betokening understanding and on one's competence with the self-concept, one is entitled to attribute that instance to oneself as oneself.

This norm is grounded in the fact that what one is entitled to believe about one's self rests fundamentally on what one is entitled to believe about one's psychological states. Selves are not systems of psychological states. They are the subjects/agents of such systems. But the natures of selves consist in their psychological competencies. As noted in the first lecture, the capacity to form warranted beliefs about one's self through beliefs about one's psychological states has its phylogenetic prehistory in couplings between occurrences of *de se*, or egocentric, indexes applied to an individual as origin of a representational framework and as subject matter of perception. Such couplings center on perceivable bodily attributes. Selves are not perceivable. But we can know our psychological states *de re*. Such knowledge is a basic route to knowing our selves. Betokening understanding of instances of psychological elements grounds the *Self*-Knowledge that I have been discussing.

A central instance of self-cognition through cognition of one's psychological states is self-tracking in memory. Again, the relevant memory is grounded in more primitive couplings. *De se*, egocentric indexes, the primitive ancestors of self-concepts, are embedded in all experiential memory.[114] When occurrences of such indexes *that mark an individual* are coupled with occurrences applied to a subject matter, they allow diachronic ego-tracking. The initial tracking is perceptual and body centered. Memories associated with couplings in passing the mirror test are likely examples. When memories become meta-psychological, indexes in them can be used in couplings that track past psychological instances, for example, transitions in reasoning, in meta-psychological autobiographical episodic memory. Such trackings are among the diachronic unities that make selves possible.

---

[114] For more on primitive ancestors of the self-concept, see my "Memory and Persons"; and "*De Se* Preservation and Personal Identity: Reply to Shoemaker," forthcoming.

These unities are grounded in present-tense couplings—those that attribute current psychological instances, identified through betokening understanding, to a subject/agent that is a central *locus* for them. Descartes' *cogito* focused on performative present-tense couplings. But there are equally important couplings with present standing states. When couplings of egocentric indexes are incorporated into betokening understanding of psychological states and used in critical reasoning, such indexes support self-concepts. Having a self-concept, indeed being a self, is grounded in reflexive psychological competencies that meet epistemic norms for Self-Understanding.

Fulfilling these epistemic norms is making connections constitutive to selves. Making the connections—engaging in the diachronic and multi-tiered preservations—just is getting things right about one's psychology and one's self. Thus fulfilling the norm set out in (WBA) yields immunity to brute error for self-attributions. The step's starting point—warranted betokening understanding—cannot introduce brute error. If self-attribution relies on that understanding, it veridically self-attributes the Self-Understood state. Self-attributions based in betokening understanding are paradigmatically warranted uses of the self-concept. Errors in self-attribution derive from warrant-undermining pathology or from not relying on the relevant preservational mechanisms. So self-attributions are immune to brute error, relative to the relevant warrants.

Conceiving the psychological elements as one's own via the self-concept is acknowledging ownership of commitments in the psychological elements. Capacity for such acknowledgment helps ground applicability of norms of critical reason and morality.

Many issues arise about this account of warrant for self-attribution. I mention two.

Shoemaker's discussion of quasi-memory may seem to threaten immunity to brute error in Self-Understanding. Through quasi-memory, an individual could identify an attitude instance that is *not* the individual's own in a *de re* way. The individual could make a brute error in self-attributing the quasi-remembered psychological state. Remembering that warrants attach to psychological abilities helps disarm such threats. An ability to preserve one's own attitudes differs, psychologically and epistemically, from any ability to tap into other minds. Brute errors of self-attribution relative to warrants associated with quasi-memory are irrelevant to self-attributions warranted by memory.[115]

---

[115] Shoemaker, "Persons and Their Pasts," *American Philosophical Quarterly*, VII, 4 (October 1970): 269–85, reprinted in his *Identity, Cause, and Mind* (New York: Oxford, 2004). See also my "Memory and Persons."

Some schizophrenics think that their thoughts are inserted into their minds by others. They deny that their thoughts, which they cite correctly, are their own. Such cases do not threaten the account because they involve pathology in the preservative capacities that ground entitlements to self-attribution. The schizophrenics do not rely purely on betokening understanding to make a self-attribution, in the way specified by the norm. There is some malfunction in connecting self-ascription with betokening understanding. Such malfunction undermines warrant for self-ascription in judgments about state instances. The errors are not errors in exercising well-functioning capacities that constitute apperceptive points of view.[116]

My account of warrant for Self-Understanding and Self-Knowledge refines Kant's notion of unity of apperception. Unity lies in an understanding that utilizes trans-temporal and inter-level constitutive connections in the structure of selves.

The account highlights a *singular intellectual* power, *de re* understanding of particulars. Descartes emphasized *de re* understanding in contrast to Aquinas's emphasis on reason as a competence with generality. Reason and *de re* intellectual understanding are deeply interrelated. But they are distinct. Even now, the latter is neglected in philosophy.

Kant took up Descartes' notion in his own conception of understanding of particulars. Kant insisted that such understanding must be warranted through sense capacities. I think the insistence mistaken. I think that Descartes was right to hold that we have an understanding of particulars that has a purely intellectual warrant. Descartes overrated the capacity's power for metaphysical insight. But he was right to highlight the capacity. Self-Understanding provides a central example. Other examples occur in mathematics, certain types of appreciation of intellectual beauty, and perhaps moral thinking.

I hope to have delineated the centrality of Self-Understanding in being a self—and in having the cognitive and valuational capacities, and being subject to the norms, that make us special. Using our intellects to understand the particular, particularly ourselves, is a way in which we are valuable, minded islands in a largely mindless, value-neutral universe.

TYLER BURGE

University of California, Los Angeles

---

[116] See Christopher D. Frith, *The Cognitive Neuropsychology of Schizophrenia* (Hove, UK: Lawrence Erlbaum, 1992); Simon Mullins and Sean A. Spence, "Re-Examining Thought Insertion: Semi-Structured Literature Review and Conceptual Analysis," *The British Journal of Psychiatry*, CLXXXII (2003): 293–98.

CAMBRIDGE

NEW *and* FORTHCOMING TITLES *in* PHILOSOPHY

Should Race Matter?
Unusual Answers to the Usual Questions
DAVID BOONIN
$99.00. Hb 978-0-521-76086-7
$34.99. Pb 978-0-521-14980-8
440 pp

The Philosophy of Human Evolution
MICHAEL RUSE
Cambridge Introductions to Philosophy and Biology
$85.00 Hb 978-0-521-11793-7
$27.99 Pb 978-0-521-13372-2
200 pp

Kant: *Natural Science*
IMMANUEL KANT
and ERIC WATKINS
The Cambridge Edition of the Works of Immanuel Kant
$130.00 Hb 978-0-521-36394-5
1,110 pp

Kant and Cosmopolitanism
The Philosophical Ideal of World Citizenship
PAULINE KLEINGELD
$90.00 Hb 978-0-521-76418-6
232 pp

Prices subject to change

Understanding Moral Obligation
Kant, Hegel, and Kierkegaard
ROBERT STERN
Modern European Philosophy
$90.00 Hb 978-1-107-01207-3
260 pp

Consciousness and the Self
New Essays
JEELOO LIU
and JOHN PERRY
$90.00 Hb 978-1-107-00075-9
272 pp

Plato and the Divided Self
RACHEL BARNEY,
TAD BRENNAN
and CHARLES BRITTAIN
$90.00 Hb 978-0-521-89966-6
320 pp

Socrates' Daimonic Art
Love for Wisdom in Four Platonic Dialogues
ELIZABETH S. BELFIORE
$99.00 Hb 978-1-107-00758-1
350 pp

Réne Descartes: *Meditations on First Philosophy*
A Latin-English Edition
JOHN COTTINGHAM
$90.00 Hb 978-0-521-19138-8
200 pp

Globalization and Global Justice
Shrinking Distance, Expanding Obligations
NICOLE HASSOUN
$85.00 Hb 978-1-107-01030-7
220 pp

Ethics and War
An Introduction
STEVEN P. LEE
Cambridge Applied Ethics
$95.00 Hb 978-0-521-89883-6
$29.99 Pb 978-0-521-72757-0
240 pp

Ethics and the Media
An Introduction
STEPHEN J. A. WARD
Cambridge Applied Ethics
$90.00 Hb 978-0-521-88964-3
$28.99 Pb 978-0-521-71816-5
298 pp

Ethical Naturalism
Current Debates
SUSANA NUCCETELLI
and GARY SEAY
$90.00 Hb 978-0-521-19242-2
248 pp

The World-Time Parallel
Tense and Modality in Logic and Metaphysics
A. A. RINI
and M. J. CRESSWELL
$95.00 Hb 978-1-107-01747-4
240 pp

Perception, Realism and the Problem of Reference
ATHANASIOS RAFTOPOULOS
and PETER MACHAMER
$95.00 Hb 978-0-521-19877-6
300 pp

Third Edition!

Chinese Philosophy
HAIMING WEN
Introductions to Chinese Culture
$19.99 Pb 978-0-521-18676-6
152 pp

www.cambridge.org/us/philosophy

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE •

from CAMBRIDGE UNIVERSITY PRESS!

Human Rights Law and the Marginalized Other
William Paul Simmons
$90.00 Hb 978-1-107-01007-9
250 pp

Cambridge History of Philosophy in the Nineteenth Century (1790–1870)
Allen W. Wood *and* Songsuk Susan Hahn
$180.00 Hb 978-0-521-77273-0
930 pp

Nietzsche's *On the Genealogy of Morality*
A Critical Guide
Simon May
Cambridge Critical Guides
$99.00 Hb 978-0-521-51880-2
360 pp

Plato's Account of Falsehood
A Study of the *Sophist*
Paolo Crivelli
$90.00 Hb 978-0-521-19913-1
220 pp

Spinoza's Geometry of Power
Valtteri Viljanen
$90.00 Hb 978-1-107-00780-2
208 pp

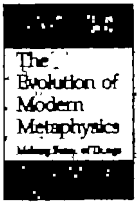
Interpreting Newton
Critical Essays
Andrew Janiak *and* Eric Schliesser
$120.00 Hb 978-0-521-76618-0
440 pp

Interpreting Suárez
Critical Essays
Daniel Schwartz
$95.00 Hb 978-0-521-50965-7
248 pp

The Political Philosophy of Zionism
Trading Jewish Words for a Hebraic Land
Eyal Chowers
$99.00 Hb 978-1-107-00594-5
220 pp

Explorations in Ancient and Modern Philosophy *by* M. F. Burnyeat
Volume 1
$130.00 Hb 978-0-521-75072-1
440 pp
Volume 2
$130.00 Hb 978-0-521-75073-8
410 pp

The Evolution of Modern Metaphysics
Making Sense of Things
Adrian Moore
The Evolution of Modern Philosophy
$99.00 Hb 978-0-521-85111-4
$39.99 Pb 978-0-521-61655-3
500 pp



Contemporary Aristotelian Metaphysics
Tuomas E. Tahko
$99.00 Hb 978-1-107-00064-3
280 pp

Kuhn's Evolutionary Social Epistemology
K. Brad Wray
$95.00 Hb 978-1-107-01223-3
244 pp

Insurrection and Intervention
The Two Faces of Sovereignty
Ned Dobos
$85.00 Hb 978-0-521-76113-0
224 pp

Epicurus and the Epicurean Tradition
Jeffrey Fish *and* Kirk R. Sanders
$90.00 Hb 978-0-521-19478-5
280 pp

Physician-Assisted Death in Perspective
Assessing the Dutch Experience
Stuart J. Youngner *and* Gerrit K. Kimsma
$99.00 Hb 978-1-107-00756-7
424 pp

Prices subject to change

www.cambridge.org/us/philosophy

CAMBRIDGE
UNIVERSITY PRESS

The Monist

An International Journal of General Philosophical Inquiry

Topics for 2012

January—Dilemmas of Multiculturalism
April—Experimental Philosophy
July—Neuroethics
October—Music

2012 *Individual* Subscription Rates:		2012 *Institutional* Subscription Rates:	
1 year (4 issues)	$45.00	1 year (4 issues)	$65.00
2 years (8 issues)	$70.00	2 years (8 issues)	$110.00

($18.00 per year will be added for postage outside the U.S.)

Online-only subscriptions to *The Monist* and its ever-expanding list of back issues may be purchased from Philosophy Documentation Center. Please visit pdcnet.org, or call 1-800-444-2419.

Back Issues: Many issues published since 1962 are available at $15.00 each. Customers within the U.S., please add $3.00 per issue for postage. Customers outside the U.S., please add $8.00 per issue for postage. Please write for discounts on larger orders or for a free list of available back issues. Express mail is available at an additional charge. Please inquire at themonist@gmail.com.

Please make payments by company or personal check or money order, in U.S. dollars, made out to *The Monist*, 315 Fifth Street, P.O. Box 599, Peru, Illinois, U.S.A., 61354-0599, or via paypal.com to the address: themonist@gmail.com. (Additional PayPal fees may apply for international and credit card payments.)

Editor: Barry Smith, University at Buffalo: phismith@buffalo.edu
Managing Editor: George A. Reisch: themonist@gmail.com

www.themonist.org

The MIT Press

Dream Life

An Experimental Memoir

J. Allan Hobson

"This remarkably candid and unconventional memoir provides a fascinating account of the life of the mind, as lived and boldly explored by one of the foremost brain scientists of our time."

— Frank J. Sulloway, University of California, Berkeley, and author of *Freud, Biologist of the Mind*

304 pp., 30 illus., $29.95 cloth

Information and Living Systems

Philosophical and Scientific Perspectives

edited by George Terzis and Robert Arp

"This volume has the virtue of airing a number of refreshing voices that are not often heard on this side of the Atlantic, and that bring perspectives that should energize our conversations about information in living systems."

— Evelyn Fox Keller, MIT

472 pp., 21 illus., $50 cloth

Modes of Creativity

Philosophical Perspectives

Irving Singer

appendix by Moreland Perkins

"The writing in this book is classic Singer: gracefully urbane, informed, insightful, and easily at home with the whole Western tradition in philosophy. Readers will welcome its wholesome sunlit sanity amid the fogs and miasmas of postmodernism."

— Thomas Alexander, Southern Illinois University Carbondale

320 pp., $36.95 cloth

The Language of Thought

A New Philosophical Direction

Susan Schneider

A philosophical refashioning of the Language of Thought approach and the related computational theory of mind.

296 pp., 2 illus., $36 cloth

Perplexities of Consciousness

Eric Schwitzgebel

A philosopher argues that we know little about our own inner lives.

A Bradford Book • 232 pp., 6 illus., $27.95 cloth

Addiction and Responsibility

edited by Jeffrey Poland and George Graham

The intertwining of addiction and responsibility in personal, philosophical, legal, research, and clinical contexts.

Philosophical Psychopathology series

328 pp., 9 illus., $40 cloth

Inside Jokes

Using Humor to Reverse-Engineer the Mind

Matthew M. Hurley, Daniel C. Dennett, and Reginald B. Adams, Jr.

"Before this book the only comedy that had been peer reviewed and replicated in double-blind experiments was the theory that there's nothing funnier than a smoking monkey. I'm so glad smart people outside of comedy are taking comedy seriously."

— Penn Jillette of "Penn & Teller"

344 pp., $29.95 cloth

To order call 800-405-1619 • **http://mitpress.mit.edu** • Visit our e-books store: **http://mitpress-ebooks.mit.edu**

Dialogues with Davidson

Acting, Interpreting, Understanding

edited by Jeff Malpas

foreword by Dagfinn Føllesdal

Leading scholars discuss Donald Davidson's work in relation to a wide range of contemporary philosophical issues and approaches

504 pp , $50 cloth

Laws, Mind, and Free Will

Steven Horst

"Start with a mix of powers, like mine or Mumford's, and Giere's perspectivism Mix in a great deal of Horst The result is a lovely pluralist view of scientific models and laws, a view that meshes neatly with the matching doctrine of cognitive pluralism that Horst devlops from his detailed case studies and that leaves ample room for—but does not force—free will"

— Nancy Cartwright, London School of Economics and University of California, San Diego

Life and Mind series • 288 pp , 60 illus , $35 cloth

Wittgenstein in Exile

James C. Klagge

"James Klagge's portrayal of Wittgenstein as an exile is original, engaging, and persuasive In a crowded field, it is a genuinely useful addition to the secondary literature and should be read by all those interested in understanding both Wittgenstein himself and his philosophy"

— Ray Monk, author of *Ludwig Wittgenstein: The Duty of Genius*

264 pp , 1 illus , $35 cloth

Game Theory and the Humanities

Bridging Two Worlds

Steven J. Brams

How game theory can offer insights into literary, historical, and philosophical texts ranging from Macbeth to Supreme Court decisions

312 pp , 35 illus , $35 cloth

Now in Paper

The Monstrosity of Christ

Paradox or Dialectic?

Slavoj Žižek and John Milbank

edited by Creston Davis

"A dazzling dialogue, not for the faint-hearted"

— Marcus Pound, *Theology*

Short Circuits series, edited by Slavoj Žižek
320 pp , $13 95 paper

Now in Paper

Things and Places

How the Mind Connects
with the World

Zenon W. Pylyshyn

"Pylyshyn is a singular figure in cognitive science—an extraordinary psychologist who is profoundly dedicated to understanding and responding to philosophical concerns *Things and Places* belongs in the bookcase of anyone who believes that uncracking philosophical puzzles about the mind requires a hefty dose of empirical study"

— Lawrence Shapiro, *Mind*

Jean Nicod Lectures • 272 pp , 26 illus , $18 paper

To order call 800-405-1619 • **http://mitpress.mit.edu** • Visit our e-books store **http://mitpress-ebooks.mit.edu**

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 8
AUGUST 2011

page
387 *Revisability and Conceptual Change in "Two Dogmas of Empiricism"* David J. Chalmers
416 *A Modest Proposal about Chance* Jenann Ismael
443 NEW BOOKS: ANTHOLOGIES

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2011 by the Journal of Philosophy, Inc

ISSN 0022-362X

UNIVERSITY OF CALCUTTA
CENTRAL LIBRARY
1404 2011

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 8, AUGUST 2011

REVISABILITY AND CONCEPTUAL CHANGE IN "TWO DOGMAS OF EMPIRICISM"

W. V. Quine's article "Two Dogmas of Empiricism" is one of the most influential works in twentieth-century philosophy.[1] The article is cast most explicitly as an argument against logical empiricists such as Carnap, arguing against the analytic/synthetic distinction that they appeal to along with their verificationism. But the article has been read much more broadly as an attack on the notion of the a priori and on the program of conceptual analysis.

I will address Quine's article construed as a critique of the notions of analyticity and apriority. I will focus especially on the most influential part of Quine's article: the arguments in the final section concerning revisability and conceptual change. In addressing these arguments, I will adopt a line of response grounded in Carnap's underappreciated article "Meaning and Synonymy in Natural Languages."[2] I will argue that an analysis inspired by this article, when conjoined with tools drawn from contemporary two-dimensional semantics and from Bayesian confirmation theory, provides what is needed to reject Quine's argument.

I will not give a positive account of the analytic or the a priori beyond the standard definition of analyticity as truth in virtue of meaning and of apriority as knowability with justification independent of experience. I am more inclined to defend the notion of apriority than the notion of analyticity, so I will focus more on the former, but the response that I will develop can be used to defend either notion from Quine's critique.

---

[1] W V. Quine, "Two Dogmas of Empiricism," *Philosophical Review*, LX, 1 (January 1951): 20–43.

[2] Rudolf Carnap, "Meaning and Synonymy in Natural Languages," *Philosophical Studies*, VI, 3 (April 1955): 33–47. Reprinted as Appendix D of Carnap's *Meaning and Necessity: A Study in Semantics and Modal Logic* (Chicago: University Press, 1956).

0022-362X/11/0808/387–415

I. THE ARGUMENTS OF "TWO DOGMAS"

In sections I through IV of "Two Dogmas," Quine argues that if one tries to make sense of the notion of analyticity, one ends up moving in a circle through cognate notions (synonymy, definition, semantic rules, meaning), and one cannot break out of the circle. Many philosophers have been unmoved by this worry, as it seems that one finds a similar circle for all sorts of philosophically important notions: consciousness, causation, freedom, value, existence. So I will set these criticisms aside here.

In section V of the article, Quine makes points that specifically address Carnap's logical empiricism, criticizing his construction of physical concepts from phenomenal concepts in the *Aufbau*, and his verification theory of meaning. I will set these points aside here, as I am not concerned to defend Carnap's construction or the verification theory of meaning.

The extraordinary influence of Quine's article can be traced in large part to its short final section. Part of this influence stems from the positive picture that Quine offers in the first paragraph of the section, characterizing the totality of our knowledge as a "man-made fabric which impinges on experience only along the edges," in which "no particular experiences are linked with any particular statements in the interior of the field, except indirectly through considerations of equilibrium, affecting the field as a whole." This picture serves as a powerful alternative to the verificationist picture provided by some logical empiricists, but it does not contain any direct argument against the analytic/synthetic distinction or the related notion of apriority.

The most influential arguments against an analytic/synthetic distinction are found in the second paragraph, which I quote in full:

> If this view is right, it is misleading to speak of the empirical content of an individual statement—especially if it be a statement at all remote from the experiential periphery of the field. Furthermore it becomes folly to seek a boundary between synthetic statements, which hold contingently on experience, and analytic statements which hold come what may. Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system. Even a statement very close to the periphery can be held true in the face of recalcitrant experience by pleading hallucination or by amending certain statements of the kind called logical laws. Conversely, by the same token, no statement is immune to revision. Revision even of the logical law of the excluded middle has been proposed as a means of simplifying quantum mechanics; and what difference is there in principle between such a shift and the shift whereby Kepler superseded Ptolemy, or Einstein Newton, or Darwin Aristotle?[3]

---

[3] Quine, *op. cit.*, p 40.

I will focus on these critical arguments. There are two crucial points.

> (Q1) "Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system."
>
> (Q2) "No statement is immune to revision."

If (Q1) and (Q2) are read as mere psychological claims, saying that as a matter of fact someone might hold onto or revise any statement, then they are highly plausible, but not much of interest will follow from them. Quine is saying something more than this. We can understand (Q1) as saying that any statement can be *rationally* held true come what may, and (Q2) as saying that that no statement is immune to *rational* revision. These points have interesting consequences.

Many have taken these points to suggest either that no sentences are analytic, or that no distinction can be drawn between analytic and synthetic sentences. One possible connection goes via the theses that analytic sentences are those that can be rationally held true come what may and that all analytic sentences are immune to rational revision. If so, (Q1) suggests that by the first criterion, all sentences will count as analytic. And (Q2) suggests that by the second criterion, no sentence will count as analytic. Either way, there is no useful distinction between analytic and synthetic sentences to be had. Similarly, if we assume that a priori sentences are those that can be rationally held true come what may and that all a priori sentences are immune to revision, (Q1) and (Q2) suggest that there is no useful distinction between a priori and a posteriori sentences to be had.

One common response to the argument from (Q2) is to suggest that revisability is quite compatible with apriority (or analyticity), on the grounds that a priori justification (or the justification we have for believing analytic sentences) is defeasible.[4] For example, I might know a mathematical claim a priori, but my justification might be defeated if I learn that a leading mathematician thinks that the claim is false. I think that this response is correct as far as it goes, but to rest entirely on it would be to concede a great deal to Quine. On a common traditional conception, at least some a priori justification (and some justification for believing analytic truths) is indefeasible.

---

[4] See, for example, Hartry Field, "The A Prioricity of Logic," *Proceedings of the Aristotelian Society*, n.s., xcvi (1996): 359–79. Philip Kitcher, "A Priori Knowledge Revisited," in Paul Boghossian and Christopher Peacocke, eds., *New Essays on the A Priori* (New York: Oxford, 2000), pp. 65–91, defends a conception of the a priori that requires indefeasibility, while Peacocke, *The Realm of Reason* (New York: Oxford, 2004) defends a conception that does not. For present purposes I will remain neutral on whether apriority entails some sort of ideal indefeasibility; the observation about testimony in section VI contains some relevant discussion.

One reasonably might hold that some a priori justification (in logic or mathematics, say) yields not just knowledge but certainty, at least on ideal reflection. These claims are not obviously correct, but they are also not obviously incorrect, and I do not think that Quine's argument establishes that they are false. So I will take another line of response.

The response I will develop takes off from the response given by Grice and Strawson at the end of their article "In Defense of a Dogma." This response holds that (Q1) and (Q2) are compatible with an analytic/synthetic distinction, for a reason quite different from the one given above. Here is a passage addressing the argument from (Q2):

> Now for the doctrine that there is no statement which is in principle immune from revision, no statement which might not be given up in the face of experience. Acceptance of this doctrine is quite consistent with adherence to the distinction between analytic and synthetic statements. Only, the adherent of *this* distinction must also insist on another; on the distinction between that kind of giving up which consists in merely admitting falsity, and that kind of giving up which involves changing or dropping a concept or set of concepts. Any form of words at one time held to express something true may, no doubt, at another time, come to be held to express something false. But it is not only philosophers who would distinguish between the case where this happens as the result of a change of opinion solely as to matters of fact, and the case where this happens at least partly as a result of a shift in the sense of the words. Where such a shift in the sense of the words is a necessary condition of the change in truth-value, then the adherent of the distinction will say that the form of words in question changes from expressing an analytic statement to expressing a synthetic statement.... And if we can make sense of this idea, then we can perfectly well preserve the distinction between the analytic and the synthetic, while conceding to Quine the revisability-in-principle of everything we say.[5]

Here the central point is that our judgments about any *sentence*, even an analytic sentence, will be revisable if the meaning of the words change. For example, if 'bachelor' changes from a term for unmarried men to a term for sociable men, then we will no longer judge that 'All bachelors are unmarried' is true. But this observation is just what an adherent of the analytic/synthetic distinction should expect. Analytic sentences instead should be understood as those sentences that are immune to revision while their meaning stays constant. More precisely, they are those that are immune to rational rejection while their meaning stays constant. (There is a sense in which an analytically false sentence might be immune to rational revision, but from here onward I will understand 'revision' as requiring rejection.)

---

[5] H. P. Grice and P. F. Strawson, "In Defense of a Dogma," *Philosophical Review*, LXV, 2 (April 1956): 141–58, at pp. 156–57.

Following standard practice, we can say that when the meaning of a sentence changes, there is *conceptual change*: some expression in the sentence at first expresses one concept and later expresses another. When the meaning of a sentence stays the same, there is *conceptual constancy*: the expressions in the sentence will express the same concepts throughout. Then Grice and Strawson's point could be put by saying that an analytic sentence is one that is immune to revision without conceptual change. More cautiously, the point could be put by saying that the fact that a sentence is revisable under conditions of conceptual change does not entail that it is not analytic. Something similar applies to apriority.[6]

At this point, Quine has two obvious replies. The first is to say that the appeal to meaning in characterizing the class of analytic sentences is circular, as the notion of meaning is as poorly understood as the notion of analyticity. The same could be said for the appeal to concepts and to propositions. This reply would be in the spirit of the first four sections of "Two Dogmas." But then this argument will not be much of an advance on the arguments in the first four sections, and anyone who is not moved by those arguments will not be moved by this one.

The second, more interesting reply is to challenge Grice and Strawson to provide a principled distinction between cases of revision that involve conceptual change and those that involve conceptual constancy. Quine might argue that cases that are purported to be on either side of this division are in fact continuous with each other, and that there is no principled distinction to be had. Something like this thought might even be read into the last sentences of the paragraph from Quine quoted above.

Now one might suggest that Grice and Strawson are not obliged to provide a reductive characterization of the distinction—that is, one that does not use 'meaning' and cognate notions—any more than they are required to provide a reductive definition of meaning or analyticity to answer the challenge in the first four sections. Again, this suggestion seems correct as far as it goes. Nevertheless, if Quine's opponent cannot say much to characterize the principled distinction here, he or she is at least in the awkward dialectical position of leaving a challenge unanswered, and of leaving doubts about the distinction unassuaged.

---

[6] We also might allow that there is conceptual change in this sense when the proposition expressed by an utterance of a sentence changes because of a shift in context. For example, 'Someone is bald iff they have no hairs' might be accepted in one context and rejected in another. It is not clear that a mere contextual shift could change the status of a sentence as analytic, as arguably the meaning of such a sentence stays constant throughout. But if we say that a sentence is a priori if it expresses a proposition that is knowable a priori, then it is natural to hold that a sentence might be a priori in one context but not in another. It is for reasons like this that I speak of 'conceptual change' rather than 'meaning change' or 'semantic change'; the latter phrases tend to suggest changes in standing linguistic meaning (thereby excluding mere contextual shifts), but it is changes in the propositions and concepts expressed that matter most for our purposes.

My view is that much can be said to flesh out a principled distinction. I think that the basic tools for doing so can be found in Carnap's "Meaning and Synonymy in Natural Languages."

II. CARNAP ON INTENSIONS

Carnap is Quine's major target in "Two Dogmas of Empiricism." It is not always appreciated that "Meaning and Synonymy in Natural Languages" can be read as a sustained response to Quine, perhaps because Carnap spends little time discussing him explicitly. Nevertheless, Carnap says enough to make clear that a response to "Two Dogmas" is intended.

Carnap's article sets out to analyze the notion of meaning and related notions such as synonymy. His aim is to provide a "scientific procedure" by which meaning and synonymy can be analyzed in broadly naturalistic terms. Importantly, he aims to explicate not only the notion of extension, but the notion of intension (the "cognitive or designative component of meaning"), which he notes has been criticized by Quine as "foggy, mysterious, and not really understandable."[7]

Carnap's key idea is that we can investigate the intension that a subject associates with an expression by investigating the subject's judgments about possible cases. To determine the intension of an expression such as 'Pferd' for a subject, we present the subject with descriptions of various logically possible cases, and we ask the subject whether he or she is willing to apply the term 'Pferd' to objects specified in these cases. If we do this for enough cases, then we can test all sorts of hypotheses about the intension of the expression.

In this article, Carnap takes the term 'intension' as primitive and does not build possible cases into the very nature of intensions. But for our purposes it is useful to adopt a suggestion that Carnap makes elsewhere, and simply define an intension as a function from possible cases to extensions. For a term like 'Pferd', the intension will be a function from possible cases to objects characterized in those cases. For a sentence such as 'Grass is green', the intension will be a function from possible cases to truth-values. Then Carnap's procedure above can be regarded as a way of directly ascertaining the values of the intension that a subject associates with an expression, by presenting the subject with a possible case and noting the extension that the subject associates with the case.

---

[7] Carnap, *op cit*, p. 36.

Of course one cannot actually present a subject with all possible cases to determine every aspect of an intension. But Carnap suggests that the intension that a speaker associates with an expression is determined by the speaker's linguistic dispositions. For a given expression *E* used by a given speaker, the speaker will have the disposition to associate a given extension with *E*, when presented with a possible case. For example, if *S* is a sentence, the speaker will have the disposition to judge the sentence as true or false of a possible case, when presented with that case. The intension of an expression can then be seen as a function that maps possible cases to the extension that the speaker is disposed to identify, when presented with that case.

In this way, Carnap defines an expression's intension in naturalistic and even operational terms. We can go on to define synonymy: two expressions are synonymous (for a speaker at a time) when they have the same intension (for that speaker at that time). And we can define analyticity: a sentence is analytic (for a speaker at a time) when its intension has the value 'true' at all possible cases (for that speaker at that time).

With this definition in hand, we can go on to provide a principled criterion for conceptual change over time. An expression *E* undergoes change in meaning between t_1 and t_2 for a speaker iff the speaker's intension for *E* at t_1 differs from the speaker's intension for *E* at t_2. If we accept Carnap's dispositional account of intensions, it follows that *E* undergoes change in meaning between t_1 and t_2 iff there is a possible case such that the speaker is disposed to associate different extensions with *E* when presented with the case at t_1 and t_2.

Of course there are many immediate questions about Carnap's account. What is a possible case? In what vocabulary are these cases specified? How can we determine whether the meaning of this vocabulary has changed? Can speakers make mistakes about intensions? Can they change their mind about a case without a change in meaning? Can meaning really be operationalized this easily? And so on. Carnap's account may need to be modified or at least refined to answer these questions.

Before addressing these matters, I will illustrate how Carnap's account might be used to address the challenge in section VI of "Two Dogmas" directly. In my view, the essential aspects, if not the specifics, of the resulting response are sound. These essential aspects carry over to more refined analyses couched in terms of two-dimensional semantics (section IV) and Bayesian confirmation theory (section V and VI).

III. A CARNAPIAN RESPONSE

In "Meaning and Synonymy in Natural Languages," Carnap does not mention the arguments in section VI of "Two Dogmas"; nor does he address revisability or conceptual change. Nevertheless, his framework can be used to give a response to these arguments that is broadly in the spirit of Grice and Strawson's response, fleshed out with a principled criterion for conceptual change.

We can start with Quine's observation that any statement can be held true come what may. This seems correct. Even a paradigmatically synthetic sentence such as 'All bachelors are untidy' can be held true in the face of apparently countervailing evidence, if we allow sufficient adjustment of ancillary claims. The question is whether such adjustments will involve conceptual change, and whether we have a principled criterion for determining this.

We might as well start with a case. At t_1, Fred asserts "All bachelors are untidy." At t_2, Fred is presented with evidence of a tidy unmarried man. Fred responds, "He's no bachelor! Bachelors must be over 30, and he's only 25." At t_3, Fred is presented with evidence of a 35-year-old man with a spotless apartment. Fred responds, "He's not tidy! Look at the mess in his sock drawer." In this way, Fred holds the sentence true throughout, and through similar maneuvers he can hold it true come what may.

Does this case involve conceptual change? We can apply Carnap's analysis to see whether Fred's intension for 'All bachelors are untidy' (call this sentence *B*) changes over the relevant time-span. Suppose that *c* is a detailed possible case in which there is an unmarried 25-year-old man with a tidy apartment. At t_2, when Fred is presented with the information that *c* obtains, he responds that 'All bachelors are untidy' is true with respect to *c*. By Carnap's criterion, Fred's intension for *B* is true with respect to *c* at t_2.

What about Fred's intension for *B* at t_1? The key question is: if Fred had been presented with a description of *c* at t_1, before he had evidence that the case was actual, would he have judged that 'All bachelors are untidy' was true with respect to *c*?

If the answer is yes, then Carnap's criterion suggests that there is no relevant conceptual change between t_1 and t_2. In this case, Fred will simply have had an unusual intension for 'bachelor' all along.

If the answer is no, then Carnap's criterion suggests that there is relevant conceptual change between t_1 and t_2. The intension of 'All bachelors are untidy' will have changed during this time, probably because the intension of 'bachelor' has changed during this time.

The same applies more generally. If a speaker's judgment concerning a case at t_2 is reflected in the speaker's dispositions to respond to such a case at t_1, we can say that the speaker's judgment concerning that case is *prefigured*. If a speaker's judgment concerning a case at t_2 is not reflected in the speaker's dispositions at t_1, we can say that the speaker's judgment concerning the case is *postfigured.* On Carnap's account, postfigured judgments involve conceptual change, but prefigured judgments do not.

In any case, we have what we need. Carnap's framework allows us to see how any sentence can be held true come what may, while at the same time allowing a principled way to distinguish between those cases of holding-true that involve conceptual change and those that do not. Something similar applies to cases of revisability, though I will not go into the details here.

IV. REFINING CARNAP'S ACCOUNT

Carnap's account of meaning is remarkably simple, and one might reasonably wonder whether such a simple account can be correct. I think that while there are problems with the account, they can be addressed in a way that preserves something of the spirit of the account, if not the letter.

Perhaps the most obvious problem is that subjects can make mistakes. A subject might miscalculate and judge that 36+27=73, and she might even be disposed to judge this to be true with respect to all possible scenarios. On Carnap's account, it will follow that '36+27=73' is analytic for the subject. But this seems the wrong result: on the face of it, the sentence is not even true. Similar mistakes seem possible for nonideal subjects in all sorts of domains.

To handle cases of this sort, we can modify the account to appeal not to what the subject would say in response to the case, but to what the subject should say, or what she would say given ideal reasoning. We might say that the intension of E maps a possible case C to the extension that the subject would identify for E, if she were to be presented with C and were to reason ideally.

Construed this way, the account will no longer yield an operational definition of meaning, at least unless we can find an operational criterion for ideal reasoning. But this is not a bad thing for those inclined to reject behaviorism in any case. It is also far from clear that this account provides a naturalistic reduction of meaning. It will do so only if we already have a naturalistic reduction of ideal reasoning. But the account need not be a naturalistic reduction to be useful.

Someone might suggest that in these cases, facts about meaning determine facts about ideal reasoning rather than vice versa: it is precisely because we mean such-and-such by 'Pferd' that we should say such-and-such. We need not take a stand on these questions about metaphysical priority here. All we need is that in these cases, there are facts about what subjects should say or what ideal reasoning dictates, and that we have some pretheoretical grip on these facts. Then we can use these facts to help explicate a corresponding notion of meaning, regardless of which of these notions is metaphysically prior. In effect, we are using an antecedent grip on normative notions to help explicate semantic notions. Of course it remains open to a Quinean opponent to reject normative notions entirely. I discuss opposition of that sort later in this article.

Another problem is that on the contemporary understanding, intensions often are inaccessible to a subject, even given ideal reasoning. For example, if Kripke is right, the intension of 'water' picks out H_2O in all possible worlds, even for subjects who do not know that water is H_2O.[8] Such subjects will not be disposed to identify H_2O as the extension of 'water' when presented with a possible case, so Carnap's definition will get the intension wrong.

To handle this problem, we can take a leaf from two-dimensional semantics, which recognizes two sorts of intension. Even in light of Kripke's insight, Carnap's account still might apply to one sort of intension, though not the other. Kripke's point applies to secondary intensions, which govern possible cases considered as counterfactual: if there had been XYZ in the oceans and lakes, water would still have been H_2O. For the purposes of Carnap's account, though, we can focus on primary intensions and stipulate that subjects consider the possible cases as actual. For example, we can ask them to suppose that XYZ actually is in the oceans and lakes in the actual world and ask for their verdict about the extension of 'water' under that supposition. Subjects plausibly will hold that 'water' picks out XYZ if that hypothesis is correct. This mirrors the familiar suggestion that the primary intension of 'water' picks out XYZ in a Twin Earth scenario, although the secondary intension of 'water' picks out H_2O there. So it is not out of the question that a Carnap-style account might work for primary intensions, which in any case are often held to be the sort of intensions that are tied most closely to apriority and analyticity.

A third issue is the nature of possible cases. For our purposes they should be akin to possible worlds. They might be centered metaphysically possible worlds (that is, worlds marked with an individual and a time), with the centering required to handle intensions for expressions such as 'I' and 'now'. They also might be regarded as epistemically possible worlds, or epistemically possible scenarios, which might be modeled by maximal consistent sets of sentences that cannot be ruled out a priori. I will not try to settle this issue here, but I will use the word 'scenario' as a generic term for the entities involved. In order to suppose that subjects can reason about entire scenarios, we can once again appeal to the idealization of what the subject would say given ideal reasoning.

---

[8] Saul Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980)

A fourth issue is the vocabulary in which a scenario is specified. Such a vocabulary will need to be rich enough that a full enough specification using this vocabulary plus ideal reasoning determines judgments about other expressions' extensions, without being so rich that a specification builds in all those expressions directly. It is a substantive claim that some such vocabulary can be found, but proponents of two-dimensional semantics have offered arguments for this claim along with suggested vocabularies.[9] The details of such a vocabulary will not matter for our purposes.

A fifth issue is the worry that subjects might change their mind about a possible case without a change of meaning. Here, one can respond by requiring, as above, that the specification of a scenario is rich enough that judgments about sentences are determined by the specification and by ideal reasoning. If so, then if subjects are given such a specification and reason ideally throughout, there will be no room for them to change their mind in this way. Changes of mind about a fully specified scenario will always involve either a failure of ideal reasoning or a change in meaning. Of course this claim requires a version of the substantive claim in the previous paragraph. I return to a version of this issue below.

The model we then reach is something like the following. The (primary) intension of an expression for a subject is a function that maps scenarios to extensions, mapping a scenario *w* to what the subject would judge to be the extension of *E* under the supposition that *w* is actual, were she ideally rational. This is not a perfect definition, but it is good enough for our purposes. This remains very much in the spirit of Carnap's definition, although the invocation of rationality makes it a normative variation on Carnap's account.

Importantly, we can use this account to provide a version of the Carnapian response to Quine's arguments given in the previous section. Conceptual change (of the relevant sort) will occur precisely when an expression's primary intension changes across time. This will happen precisely when the subject's dispositions to judge the expression's extension in a possible case (given ideal reasoning) changes. As in the last section, we can find cases of holding-true where the dispositions change in this way, and cases where they do not. What matters is that we have a principled distinction.

---

[9] See, for example, David J. Chalmers and Frank Jackson, "Conceptual Analysis and Reductive Explanation," *Philosophical Review*, cx, 3 (July 2001): 315–60. See also my *Constructing the World* (New York: Oxford, forthcoming).

A residual issue concerns the meaning of the basic vocabulary. If cases are specified in this vocabulary, then we need to ensure that the basic vocabulary does not change in meaning throughout the process. If we do not require this, the resulting condition for meaning change will be inadequate; a subject's dispositions to judge that *S* obtains with respect to a case specified by *D* might change over time, not because the meaning of *S* changes but because the meaning of terms in *D* change. If we do require this, however, it appears that we need some further criterion for meaning change in the basic vocabulary items used in *D*, as the dispositional method would yield trivial results. So it appears that the dispositional method for determining meaning change, even when idealized, is incomplete.[10]

A second residual issue concerns the role of the a priori in characterizing this account. It is natural to suggest that the ideal reasoning in question must be restricted to ideal a priori reasoning. In fact, some two-dimensional accounts use the notion of apriority in defining primary intensions: the primary intension of a sentence *S* evaluated at a world *w* is true precisely if a material conditional 'If *D*, then *S*' is a priori, where *D* is a canonical specification of *S*.[11] If we take this route, then we have arrived at a principled distinction only by helping ourselves to the contested notion of apriority along the way.

As before, it is not clear how bad these residual problems are. One still might see the intensional analysis as demonstrating that the Quinean phenomena of holding-true and revisability are quite compatible with the intensional framework and have no power to refute it. Even if one has to assume some independent grip on the notion of apriority and on the meaning of expressions in the basic vocabulary, one can use the framework to provide a reasonably enlightening analysis of relevant cases. Still, we have not broken out of the Quinean circle. It would be nice to be able to characterize the relevant distinctions without such a direct appeal to the contested notions.

---

[10] This objection is related to Quine's argument from the indeterminacy of translation in *Word and Object* (Cambridge: MIT, 1960). Quine took Carnap's account to be a serious challenge to his arguments in "Two Dogmas," and the indeterminacy argument can be seen in part as a response to it. There, Quine argues that no dispositional analysis can settle facts about meaning, because multiple assignments of reference will always be compatible with a subject's behavioral dispositions. This applies even to Carnap's account, if we allow multiple potential assignments of reference to the basic vocabulary. In effect, Carnap's account assumes that the meaning of the basic vocabulary is fixed, but it is not clear why such an assumption is legitimate, and it is not clear how this meaning itself might be grounded in dispositional facts. Thanks to Gillian Russell for discussion here.

[11] See, for example, Chalmers, "Epistemic Two-Dimensional Semantics," *Philosophical Studies*, cxviii, 1/2 (March 2004): 153–226.

I think that such a characterization can be found. The key idea is to cast things in terms of conditional probability, rather than in terms of apriority.

An initial observation is that something very much like a primary intension can be characterized without appealing to apriority, by appealing to conditional probability instead. In particular, one can define the intension of a sentence S at a scenario w, for a subject, in terms of the subject's rational conditional credence in S given D $cr'(S|D)$, where D is a canonical specification of w. We can say that the intension of S is true at w iff $cr'(S|D)$ is high, and false at w iff $cr'(S|D)$ is low. Here we require an idealization, so that $cr'(S|D)$ is the conditional credence that the subject would have given ideal reasoning, or something along those lines.

If we do this, then we will have a principled criterion for conceptual change that does not appeal to apriority. On this criterion, a subject's intension for S will change between t_1 and t_2 iff there is a scenario w with canonical specification D such that $cr'(S|D)$ changes from high to low or vice versa. One then could run the arguments of the previous section again using this notion. This will provide a reply to Quine's challenge that gets around the second residual issue above (regarding apriority), though it still may be subject to a version of the first issue (regarding the basic vocabulary).

At this point, however, I think an alternative analysis involving conditional probability is available. This analysis is closely related to the one just mentioned, and is a descendant of the Carnapian analysis in the previous section, but it does not require any of the semantic apparatus used in that section and this one. This analysis proceeds using only standard Bayesian considerations about evidence and updating. In addition to the advantage of familiarity, this approach has other significant advantages in responding to Quine's challenge. By avoiding the need for canonical specifications of complete possible scenarios, it avoids the large idealization needed to handle enormous specifications. It also has the potential to avoid or minimize both residual issues above.

V. A BAYESIAN ANALYSIS OF HOLDING-TRUE

In what follows I develop a Bayesian analysis of Quine's arguments from holding-true and from revisability. For the purposes of this analysis, we can set the framework of scenarios and intensions to one side. All we need are orthodox Bayesian claims about credence and its revision in light of new evidence.

Let us assume a standard Bayesian model on which sentences are associated with unconditional and conditional credences for subjects at times. That is, for a given subject and a given time, a sentence S will be associated with an unconditional credence $cr(S)$, and a pair of sentences S and T will be associated with conditional credence $cr(S|T)$. (These ordinary credences $cr(S|T)$ should be distinguished from the idealized rational credences $cr'(S|T)$ from the previous section.) Credences are standardly taken to be real numbers between 0 and 1, but for our purposes exactitude is not required. It is enough that some credences be high and others low.

I will also assume a version of the standard Bayesian principle of conditionalization: if a subject has credence $cr_1(S|E)$ at t_1 and acquires total evidence specified by the evidence sentence E between t_1 and t_2, then the subject's credence $cr_2(S)$ at t_2 should be equal to $cr_1(S|E)$. I will give a more precise version of this principle below. The nature of evidence sentences will be discussed later in this article, but for now we can think of them as specifying either that certain experiences obtain or that certain observable states of affairs obtain.[12]

We can start with a typical case whereby an apparently synthetic sentence is held true in the face of apparently countervailing evidence, by appeal to appropriate ancillary theses. As before, suppose that at t_1 Fred asserts, "All bachelors are untidy." At t_2, Fred acquires evidence indicating that there is a tidy, unmarried 25-year-old man, and responds by denying that the man is a bachelor, as bachelors must be over 30.

Let B be 'All bachelors are untidy', and let E be Fred's total relevant evidence acquired between t_1 and t_2. Let $cr_1(B)$ and $cr_2(B)$ be Fred's credences in B at t_1 and t_2 respectively. Then $cr_1(B)$ and $cr_2(B)$ are both high.

The crucial question is: What is $cr_1(B|E)$, Fred's conditional credence in B given E at t_1, before Fred acquires the evidence in question?

If $cr_1(B|E)$ is high, then Fred's judgment at t_2 reflects a conditional credence that he already had at t_1. In this case, the judgment at t_2 is *prefigured*, in a sense analogous to the sense discussed earlier. Here, Fred's accepting B in light of E accords with the principle of conditionalization.

If $cr_1(B|E)$ is low, then Fred's judgment at t_2 fails to reflect the conditional credence that he already had at t_1. In this sort of case, the judgment at t_2 is *postfigured*, in a sense analogous to the sense discussed earlier. Here, Fred's accepting B in light of E appears to violate the principle of conditionalization.

---

[12] The arguments I present here also can be run using the principle of Jeffrey conditionalization, which allows conditionalization on evidence about which a subject is not certain. See Richard C. Jeffrey, *The Logic of Decision*, 2nd ed. (Chicago: University Press, 1983).

On standard Bayesian assumptions, there are two central ways one can obtain apparent violations of conditionalization for sentences. First, this can happen when the subject is not fully rational throughout the process; perhaps at t_1 Fred has not thought things through properly, or at t_2 he makes some sort of reasoning error. Second, the content of the key sentence B can change between t_1 and t_2. This can happen in cases involving indexicals, which are not relevant here, or in cases of conceptual change. In these cases, it remains possible that Fred's credences in relevant propositions obey conditionalization, but that his credences in associated sentences do not, because the association between sentences and propositions changes over time.[13]

We might formulate this as a version of the Bayesian principle of conditionalization, for sentences:

> (CS) If a subject is fully rational, and if the subject acquires total evidence specified by E between t_1 and t_2, and if the content of sentence S does not change between t_1 and t_2, then $cr_2(S) = cr_1(S|E)$.

Perhaps the most familiar version of the principle of conditionalization is cast in terms of propositions: if a fully rational subject acquires total evidence specified by proposition e between t_1 and t_2, then $cr_2(p)=cr_1(p|e)$. (CS) follows from this claim in conjunction with the plausible claims that when sentence S expresses proposition p for a subject at that time, $cr(S) = cr(p)$ at that time, and that the content of a sentence is the proposition it expresses.

It follows that if Fred in the postfigured case above is fully rational, then this is a case of conceptual change. Of course it might be that Fred is not fully rational, but this is of no help for Quine. It is unremarkable that irrational subjects might hold on to any sentence or reject any sentence, and this observation has no consequences regarding analyticity or apriority. For Quine's observations about revisability and holding-true to have any bite, rational subjects are required. So we may as well assume that Fred is fully rational.

---

[13] A potential third way that conditionalization can be violated arises on views where sentences express certain relativistic contents. Consider a view on which utterances of the sentence 'It is raining' always express the same temporal proposition *It is raining*, which can be true at some times and not at others. On Saturday, I might have a low conditional credence in *It is raining* given *The weather forecast says rain on Sunday*, then on Sunday I might acquire evidence that the weather forecast says rain on Sunday, resulting in high credence in *It is raining*, without irrationality. On a more standard view, on which the content of 'It is raining' uttered at t is *It is raining at t*, this will be classified as a change in content, but on the temporal view the content stays the same. For present purposes, we can either count these as changes in content in an extended sense, or we can require in principle (CS) that the content in question is nonrelativistic content.

If we assume that the relevant subjects are fully rational, we now have a principled criterion for conceptual change in a case of holding-true. Suppose that our subject accepts S at t_1, acquires apparently countervailing evidence E between t_1 and t_2, and continues to accept S at t_2. Then we can say

(i) If $cr_1(S|E)$ is low, this is a case of conceptual change.
(ii) If $cr_1(S|E)$ is high, this need not be a case of conceptual change.

One can now ask: is it true that a subject can hold on to any given sentence S come what may, in light of any evidence, without irrationality or conceptual change? By this analysis, this claim requires that for any given sentence S and any evidence E, $cr(S|E)$ is high (or at least is not low). But this claim is obviously false. For rational subjects and most sentences (including most paradigmatically empirical sentences), there will be evidence sentences E such that $cr(S|E)$ is low.

The moral here is that in the general case, Quinean holding-true come what may requires widespread violation of conditionalization, which requires irrationality or conceptual change. But the fact that an irrational subject might reject a sentence is no evidence that it is not analytic or a priori,[14] and the fact that a subject might reject a sentence after conceptual change is no evidence that it is not originally analytic or a priori. So Quine's argument from holding-true fails.

VI. A BAYESIAN ANALYSIS OF REVISABILITY

For our central example of revisability, we can use a familiar case from Hilary Putnam.[15] Let C be 'All cats are animals'. This might seem paradigmatically analytic or a priori. But let E specify evidence confirming that that the furry, apparently feline creatures that inhabit our houses are actually remote-controlled robots from Mars, while all the other creatures we see are organic. Putnam argues that if we discovered that E obtains, we would reject C. So let us suppose that Sarah accepts C at t_1, acquires total evidence as specified by E, and rejects C at t_2.

Here, the diagnostic question is: What is Sarah's initial conditional probability $cr_1(C|E)$?

If $cr_1(C|E)$ is low, then Sarah's judgment at t_2 reflects a conditional credence that she had at t_1. In this case, the judgment at

---

[14] Perhaps there are certain strong conceptions of analyticity on which an analytic sentence cannot be rejected by any subject, rational or irrational. But these conceptions are not standard, and in any case no such constraint applies to apriority.

[15] Hilary Putnam, "It Ain't Necessarily So," this JOURNAL, LIX, 22 (Oct. 25, 1962): 658–71.

t_2 is prefigured. Here, Sarah's accepting C in light of E accords with the principle of conditionalization.

If $cr_2(C|E)$ is high, then Sarah's judgment at t_2 fails to reflect the conditional credence that she had at t_1. In this sort of case, the judgment at t_2 is postfigured. Here, Sarah's accepting C in light of E appears to violate the principle of conditionalization.

For exactly the reasons given before, the postfigured case requires either that Sarah is not fully rational or that her use of C undergoes conceptual change between t_1 and t_2. Cases of this sort are of no help to Quine. Again, the fact that an irrational subject might reject a sentence is no evidence that it is not analytic or a priori, and the fact that a subject might reject a sentence after conceptual change is no evidence that it is not originally analytic or a priori.

For Quine's argument to succeed, he needs to exclude cases of this sort. That is, he needs to make the case that any sentence can be rationally revised without a violation of conditionalization. This requires that for all rational subjects and for all sentences S, there exists an evidence sentence E such that $cr(S|E)$ is low.

This claim is not so obviously false as the corresponding claim about holding-true come what may. For this reason, one might regard the argument from revisability as a stronger argument than the argument from holding-true. Indeed, supporters of Quine such as Putnam and Harman have concentrated on the argument from revisability and made claims not far from the claim in question.[16]

Still, it is not clear just what the grounds are for accepting the key claim. At this point, a number of observations can be made.

First, Quine's official grounds for the revisability claim involve the ability to revise ancillary claims when necessary. These grounds are the same as for the holding-true claim, and it is clear that Quine sees the two as continuous. These grounds suggest that after obtaining evidence, a subject could use these features to revise a given sentence. But we have seen that revisions of this sort typically involve violations of conditionalization. These grounds do very little to suggest that before acquiring the relevant evidence, a subject's conditional credence $cr(C|E)$ will be low.

---

[16] See Putnam, *op. cit.*; and Gilbert Harman, "Doubts about Conceptual Analysis," in Michaelis Michael and John O'Leary-Hawthorne, eds., *Philosophy in Mind: The Place of Philosophy in the Study of Mind* (Boston. Kluwer, 1994), pp 43–48. It may be useful to distinguish a pragmatist reading of the arguments in "Two Dogmas," which stresses the freedom to adjust ancillary hypotheses as one chooses, from an empiricist reading, which stresses the role of unexpected evidence in driving us to revise our beliefs. Roughly, where the pragmatist reading turns on the claim that one *may* accept or reject certain statements, the empiricist reading turns on the claim that one *should* (or perhaps that one *would*). A pragmatist reading will put equal weight on the argument from holding-true and revisability, while an empiricist reading will put more weight on the latter. The pragmatist strand is more central in Quine's text, but the empiricist strand has been more influential among later Quineans

Second, almost any claim could be rationally rejected given testimony of an apparent epistemic superior. But this point has no bearing on apriority; that a claim could be rejected in this way is no evidence that it is not a priori. The point also does not establish that any claim is revisable under ideal reflection, as it is far from clear that this sort of revisability applies to ideally rational thinkers. Perhaps these thinkers' grounds for accepting a mathematical claim, for example, will always defeat any evidence concerning an apparent epistemic superior.

Third, even if this sort of consideration applies to many apparent cases of a priori truths, there are a number against which it has no purchase. Some such cases include the material conditionals discussed in the previous section: 'If D, then S' (equivalently, '$D\rightarrow S$'), where D is a lengthy specification of an arbitrary scenario and S is a sentence such as 'Water is H_2O' such that $cr(S|D)$ is high. Assuming a fully rational subject, it follows that $cr(D\rightarrow S|D)$ is high, so that $cr(D\rightarrow S)$ is also high. We can stipulate that D includes or entails a full specification of evidence that obtains in the scenario, so that D entails E for any evidence sentence E that obtains in the scenario, and that D entails $\sim E$ otherwise (setting vagueness aside).[17] A quick two-case argument now suggests that no evidence E could lead us to rationally reject $D\rightarrow S$. First case: if E does not obtain in the scenario, then D entails $\sim E$. In this case, $cr(\sim D|E) = 1$, so $cr(D\rightarrow S|E) = 1$. Second case: if E obtains in the scenario, then D entails E. Now $cr(D\rightarrow S|E)$ must lie between $cr(D\rightarrow S|E\&\sim D)$ and $cr(D\rightarrow S|E\&D)$. But the former is 1, and the latter is just $cr(D\rightarrow S|D)$, which we have seen is high. So $cr(D\rightarrow S|E)$ is high. Putting the two cases together, $cr(D\rightarrow S|E)$ is high for all E. Importantly, material conditionals very much like these are the a priori truths that are most important in the two-dimensional framework.

Fourth, once one notes that this argument allows some truths S such that $cr(S|E)$ is high for all E, it is clear that there is no longer a sound principled argument that for all S, there is an E such that $cr(S|E)$ is low. As a result, we may expect to find many more exceptions to this claim. Indeed, many Quineans have conceded such exceptions, especially in the domains of mathematics and logic, and there is no reason not to expect many more.

---

[17] One can understand entailment here in a variety of ways. For present purposes we need only the claim that if D entails E, then $cr(E|D) = 1$ and $cr(\sim D|\sim E) = 1$ for a fully rational subject.

Fifth, it is worth stressing that even if this line of argument succeeded, it would be much more conservative than Quine's original line. It leads naturally to a view on which there is an analytic/synthetic distinction. At worst, it would be the case that most or all sentences previously regarded as analytic (a priori), such as 'All cats are animals', will be reconstrued as synthetic (a posteriori).[18] But one could still use the current framework to characterize Carnapian intensions, once one acknowledges that the intensions for sentences such as 'All cats are animals' will be false at some scenarios. One will still have a principled distinction between cases that involve conceptual change and cases that do not. In this way, advocates of analyticity, apriority, and conceptual analysis will have much of what they want.

In any case, the Bayesian analysis has given us what we wanted: a principled criterion for identifying cases of conceptual change. It has only given us a sufficient condition, rather than a necessary and sufficient condition, but this is good enough for our purposes. With this analysis in hand, it is clear that Quine's arguments from revisability and holding-true fail.

VII. QUINEAN OBJECTIONS

Objection 1: The Bayesian Analysis Begs the Question. It might be suggested that the Bayesian principle (CS) that I have appealed to simply assumes a notion of conceptual change without argument, and therefore begs the question against the Quinean skeptic about this notion. I do not think that this is quite right. (CS) is itself a consequence of the principle of conditionalization for propositions and of two other weak assumptions, none of which say anything about conceptual change. Still, this line of argument assumes a notion of proposition, about which a Quinean might be skeptical.

Now, Quine's doubts about propositions have been much less influential than his doubts about the analytic/synthetic distinction. It is clear that to get off the ground, Bayesian accounts of confirmation require either something like propositions or something like the notion of conceptual change. Bayesian credences will be assigned either to abstract entities such as propositions, events, or sets; to linguistic items such as sentences; or to mental items such as beliefs.

---

[18] Here we can distinguish between radical Quineans, who hold that there is no analytic/synthetic distinction, and moderate Quineans, who hold that there is such a distinction but that very few sentences are analytic. If the analysis given here is right, we might expect this distinction to correlate to some degree with the distinction between pragmatist and empiricist Quineans mentioned in footnote 16. This seems to be what we find in practice. For example, Quine in "Two Dogmas" takes the pragmatist and radical lines, while Putnam, *op. cit.*, takes the empiricist and moderate lines.

If we take the first route, then we can use these entities just as we used propositions to ground a notion of conceptual change. If we take the second or the third route, we need to require something like conceptual constancy to avoid counterexamples to principles such as conditionalization.

Of course a Quinean might simply reject Bayesianism altogether, along with the associated principle of conditionalization. This would seem rash, however, as Bayesianism is an extremely successful theory with widespread empirical applications. So by a Quinean's own lights, it is hard to reject it. Furthermore, even if one rejects Bayesianism, a successor theory is likely to have corresponding principles of diachronic rationality governing how beliefs should be updated over time in response to evidence. Precisely the same issues will arise for these principles: if they apply to abstract items we can use these to define conceptual change, and if they apply to linguistic items or mental items we will require a notion of conceptual change.[19]

I think the deeper moral is that there is a constitutive link between rational inference and conceptual constancy. Issues such as those floated here will arise for any principle of diachronic rationality. If it is a principle that from A and $A \rightarrow B$ one should infer B, and if the premises and conclusions here are sentences or mental items, then to avoid obvious counterexamples the principle should require that A and B have the same meaning on each occasion when they occur. And if the principle applies to abstract objects such as propositions, these can themselves be used to define conceptual change. So if we are not skeptics about principles of diachronic rationality, a notion of conceptual change will be hard to avoid.

Objection 2: Rationality Presupposes Apriority. It might be suggested that in appealing to the notion of rationality, the notion of apriority is smuggled in. For example, someone might hold that all principles of rational inference depend on underlying principles about the a priori; for example, perhaps an inference from some premises to a conclusion is rational precisely if it is a priori that if the premises obtain, the conclusion is likely to obtain. Or perhaps the distinctive idealization made by the Bayesian involves some tacit assumptions about the a priori. For example, perhaps the Bayesian requirement that rational subjects should have credence 1 in logical truths depends in some way on the belief that logical truths are a priori. If so, the appeal to rational principles here presupposes one of the key notions at issue.

---

[19] A related objection is that the very idea of a credence or a conditional credence presupposes conceptual constancy. After all, one's credence associated with a sentence arguably is determined by one's dispositions to make certain judgments and decisions involving the sentence, for example, the odds one would take on a bet on that sentence if it were offered. But in considering such dispositions, we have to assume that the meaning of the sentence stays constant from the initial moment to the bet. Likewise, a conditional credence $cr(M|E)$ arguably is determined by one's dispositions to make judgments about M conditional on the supposition of E. This requires conceptual constancy: if meaning changes between the initial moment and the judgment, a high initial credence might go with a negative judgment. Still, any conceptual constancy needed here is at best very local, within an episode of consciousness. In any case, if it turns out that the notion of apriority is as secure as the notion of a credence, so that the Quinean can reject the former only by rejecting the latter, that should be good enough for the defender of apriority. Thanks to Ned Block and Kelvin McQueen for discussion here.

The reply here is straightforward. Whether or not the objector is correct that rationality depends in some way on apriority, the appeal to rationality is innocuous in the current dialectical context. The relevant opponents are those who accept the notion of rationality but who question the notion of apriority. My argument is intended to establish that *if* one accepts certain principles concerning rationality, then one should reject Quine's argument against the a priori. If this objector is correct, then the opponent either should give up on the principles concerning rationality or accept the notion of the a priori. Either outcome is sufficient for my purposes. I am happy to concede that if an opponent rejects the notion of rationality, or rejects all relevant principles of diachronic rationality, then the current argument has no purchase against her.

It is also worth noting that the principles of rationality that I appeal to are principles that many or most opponents of the a priori accept. Conditionalization has no obvious connection to the a priori, for example. I do not know whether the special status that the Bayesian gives to logical truths has a special connection to the a priori, but in any case this status plays no role in my argument. That is, the argument does not require the Bayesian claim that rationality requires credence 1 in logical truths. In fact, the picture I have sketched appears to be compatible with a view on which logical truths deserve rational credence less than 1, and on which they can be revised given relevant evidence. All that is required is that such a revision should obey conditionalization. Nothing here smuggles in any obvious presuppositions about the a priori.[20]

Objection 3: A Principled Line between Conceptual Change and Irrationality Cannot Be Drawn. A Quinean may suggest that our concept of rationality is not fully determinate, and that as a result a clear division between cases of irrationality and cases of conceptual change cannot be found. Some hard cases, such as revising logic in light of quantum mechanics, are not easily classified as either.

---

[20] Of course, my argument appeals to logical claims at various points, but this does not require that logical truths are a priori or that they are unrevisable. It merely requires that they are true. Likewise, my argument does not require that the principle of conditionalization is a priori or that it is unrevisable. It simply requires that the principle is true.

However, my reply to Quine's argument does not require drawing a line here. It suffices for the purposes of the argument that the violations of conditionalization involve either irrationality or conceptual change, and we do not have to classify these violations further. In any case, as long as there are clear cases of rational judgment, the existence of unclear cases entails at worst a vague distinction, not a nonexistent distinction.

Objection 4: The Argument Requires Constancy in Evidence Sentences. Recall the first residual issue for the framework of intensions discussed earlier: the framework assumes conceptual constancy in the base vocabulary, so the framework cannot explain this conceptual constancy. One might think that an analogous issue arises with respect to the evidential vocabulary used to specify evidence sentences such as E. After all, conditionalization concerns what to do when one has a certain credence $cr(S|E)$ and then learns E. The conditional credence is in part an attitude to a sentence E, and what one learns is also a sentence E. One might think it is required that the sentence have the same meaning on both occasions. If so, then any apparent failures of conditionalization in a rational subject could be blamed on a change in the meaning of terms in E, instead of a change in the meaning of terms in S, and it is not clear that we have a principled way to choose.

As it stands, this picture is not quite right. Learning E does not typically involve the *sentence* E at all. Perhaps if learning were always by testimony, and if E were a sentence used in testimony, then the issue would arise. But for our purposes we can assume that the relevant learning is by perception. Here, E will be a sentence characterizing the evidence that one learns, and the learning process need not involve this sentence at all. So there is no use of E at t_2 that needs to be aligned with the use of E at t_1. At best we need to require that E as used at t_1 correctly applies to the evidence acquired at t_2. But this is a much weaker requirement, concerning only the extension of E as used at t_1, with no role for any use of E at t_2.

Still, it can be argued that acquiring evidence requires having certain *attitudes* to the evidence. For example, the rationality of Bayesian conditionalization on new experiences arguably requires not just that one has the experiences, but that one is certain that one has them. If so, one might suggest that the framework tacitly requires that at t_2, one is certain of the evidence statement E (which says that certain experiences obtain). This issue is starker in alternative frameworks such as Jeffrey conditionalization, which accommodate uncertainty about evidence by giving an explicit role to one's credence in evidence statements such as E at t_2. Does this require some sort of constancy in the meaning of E after all?

The issue is delicate. For the reasons given above, I think that the sentence E as used at t_2 plays no essential role here. However, it is arguable that at t_2, the subject must be certain of (or have other appropriate attitudes to) the evidential proposition expressed by E at t_1: the proposition that certain experiences obtain, for example. Or without invoking propositions: subjects must be certain that the relevant evidence obtains (that they are having certain experiences, say), where this is the same evidence concerning which they had conditional credences at t_1. Without this alignment, one could always respond to an apparent failure of conditionalization by saying that although a subject's initial credence was conditional on evidence e obtaining, and although evidence e later obtained, the subject in fact became certain that some other evidence e^* obtained. If this were so, there would be no violation of conditionalization (the subject would not acquire the evidence e), and there arguably would be no irrationality.

This requirement of alignment provides some room for the Quinean to maneuver, but the room is extremely limited. To eliminate this room altogether, we need only suppose that we have a grip on what it is for a subject to accept or suppose that certain evidence obtains. With this much granted, we can simply stipulate that for our purposes, the conditional credences $cr(S|E)$ relevant at t_1 are credences in S conditional on the evidence that is actually obtained at t_2. This removes any loophole, and does so without making any assumptions about constancy in the meaning of language across time. At most, we have to assume an understanding of certain beliefs and suppositions about evidence.

The required assumptions can be reduced further by noting that for our purposes, evidence can be limited to experiences, or at least to observational states of affairs. While there is a sense in which empirical knowledge of nonobservational states of affairs can serve as evidence for other claims, it is plausible that this knowledge is grounded in evidence concerning experiential or observational matters. On a Bayesian view, our credences in these states of affairs then must match those determined by conditionalization on experiential or observational matters. I think it is also plausible that credences in observational states of affairs should match those determined by conditionalization on experiential matters.[21] If the latter claim is granted, then for present purposes we can restrict the relevant evidence in cases of revisability and holding-true to experiential states. Even without it, we can restrict the relevant evidence to observational states. So to answer the Quinean worry, we need only suppose that we have a grip on what it is for a subject to accept or suppose that certain experiential or observational states of affairs obtain. This is something that Quine's arguments in "Two Dogmas" do not give us reason to doubt.

The upshot of all this is that the residual issues about a base vocabulary are not eliminated altogether on a Bayesian approach, but they are minimized in a way that brings out the severe costs of the Quinean position. A Quinean who rejects the notions of analyticity and apriority along present lines also must insist that there is no objective fact of the matter about whether a subject accepts or supposes that a given observational state obtains. This view goes along with a generalized skepticism about the contents of thought, perhaps in the spirit of Quine's skepticism about meaning developed in his arguments concerning radical translation. It likewise requires a certain skepticism about diachronic rationality, for reasons discussed earlier.

Quine himself argues both for skepticism about meaning (in *Word and Object*) and for a sort of skepticism about norms of rationality (in "Epistemology Naturalized"[22]). Few have been prepared to follow him here, and even those who sympathize with the Quine of "Two Dogmas" have tended to reject these later views. Of course Quine's arguments for these views deserve attention in their own right, but it is clear that the arguments in "Two Dogmas" provide little direct support for them. Still, the current analysis suggests a deep linkage between these views. Defending the arguments of "Two Dogmas" against a certain appeal to conceptual change leads naturally to skepticism about diachronic rational principles and about the content of language and thought. Contrapositively, once even minimal claims about rationality and thought are accepted, the arguments I have considered against analyticity and apriority dissolve.

---

[21] For example, if one is fully rational, one's credence that there is a red square in front of one should match one's antecedent conditional credence that there is a red square in front of one given that one is having an experience as of a red square. (If norms of rationality do not ensure certainty about the experiences one is having, one can move to a Jeffrey-conditionalization analog.) Theses of this sort have been denied by some dogmatists about perception (for example, James Pryor, "Uncertainty and Undermining," URL: www.jimpryor.net/research/papers/Uncertainty.pdf). They also might be denied by some who think that perceptual knowledge is more secure than introspective knowledge (for example, Eric Schwitzgebel, "The Unreliability of Naive Introspection," *Philosophical Review*, cxvii, 2 (April 2008): 245–73).

[22] Quine, "Epistemology Naturalized," in *Ontological Relativity and Other Essays* (New York: Columbia, 1969)

Objection 5: There Can Be Rational Revision by Resetting Priors. Quineans of a pragmatist stripe often appeal to the underdetermination of theory by evidence: multiple theories are consistent with the same evidence, and we have considerable latitude in choosing between them. In the Bayesian framework, where theory is determined by evidence along with prior probabilities, this underdetermination comes to underdetermination of probabilities that are prior to any evidence. This underdetermination yields a potential way that fully rational subjects might violate conditionalization without conceptual change.

The relevant method here is that of resetting priors. This method stems from the observation that most Bayesians allow that there is some flexibility in one's ultimate priors: the prior probabilities that a subject should have before acquiring any empirical evidence. (Of course these priors are something of a fiction.) For example, on Carnap's framework for inductive logic, equally rational subjects may have different values for λ, the parameter that guides how quickly the subjects adjust their beliefs in light of inductive evidence, and this difference can be traced to a difference in ultimate priors. Two such subjects might acquire exactly the same evidence over time, while being led to quite different posterior probabilities. If *G* is the thesis that human-caused global warming is occurring, for example, one subject might be led to a high credence in *G*, while another might be led to a low credence in *G*.

Now, a subject with a high credence in *G* might reflect and observe that her high credence is traceable entirely to the value of λ in her ultimate priors, and that this value was quite arbitrary. She may note that it would have been equally rational to start with a lower value of λ and to end up with a lower credence in *G*. At this point, a bold subject might choose to change her credences wholesale. At least if she has a good enough record of her evidence, she can "unwind" back to the ultimate priors, reset l to a lower value, and reintegrate all the evidence by conditionalization. She will end up with a new set of credences, including (among many other differences) a lower value for *G*.

A Quinean might suggest that there is nothing irrational about doing this, and that this method might be exploited in order that a subject could hold on to almost any sentence come what may and also revise almost any sentence. After all, for most nonobservational empirical sentences *S* and most paths of evidence, there is some ultimate prior that will lead to a high credence in *S* and some ultimate prior that will lead to a low credence in *S*. None of this requires conceptual change. So violations of conditionalization in a rational subject do not provide a sufficient condition for conceptual change after all.

This position requires rejection, or at least revision, of orthodox Bayesianism. On the orthodox view, conditionalization is a constraint on diachronic rationality, and this sort of revision will be irrational. Furthermore, the view tends to lead to an anything-goes view of rational belief. If there are no constraints on ultimate priors, the view entails that at any moment, if $cr(p) < 1$, then one's credence can be rationally revised so that $cr(p)$ is arbitrarily close to zero. And even if there are constraints on ultimate priors, these constraints must be weak enough to vindicate the large violations of conditionalization that the Quinean argument requires, leading naturally to a view on which most beliefs can be rationally revised at any moment into disbelief. Given this much, it is not easy to see how beliefs can constitute knowledge at all.[23]

Further, it is far from clear that all beliefs can be revised in this way. For example, given that logical beliefs, mathematical beliefs, and evidential statements are constrained to have credence 1, this method will not yield revisability for these beliefs. More generally, there is not much reason to hold that it will yield revisions to those beliefs usually classified as a priori ('All bachelors are unmarried', say), most of which appear not to depend on ultimate priors. So this response is weakest where it needs to be strongest.

Most fundamentally: as long as we have a conceptual distinction between cases in which beliefs are revised by this process and cases in which they are not, we still have enough to draw a distinction between those violations of conditionalization that involve conceptual change and those that do not. The Quinean will have to insist that we do not have a grip on this conceptual distinction, so that there is no distinction to be drawn between cases of resetting priors and cases of conceptual change. I think there is little reason to accept this. Furthermore, even if this line were accepted, it would lead once again to an across-the-board skepticism about principles of belief updating and other forms of diachronic rationality. So if principles of diachronic rationality are allowed at all—even the liberal principles suggested by the current approach—then the distinction between conceptual constancy and conceptual change remains intact.

---

[23] In addition, this method is a sort of belief revision that is not driven by evidence at all. So although this line of reasoning is perhaps the best way of preserving the pragmatist reading of Quine's arguments in light of the present analysis, it does not sit easily with the more influential empiricist reading.

Objection 6: Subjects Need Not Have Conditional Credences. It might be objected that the Bayesian analysis requires the assumption that for every sentence S used by a subject and every possible evidence sentence E, the subject has a conditional credence $cr(S|E)$. But this is an unrealistic idealizing assumption.

In response: the idealization is not enormous. For most S and most E, the subject will have some relevant dispositions involving S and E, for example, involving her willingness to accept various bets involving S and E. In many cases, these dispositions will line up in a clear enough way that $cr(S|E)$ will be high. In other cases, they will line up in a clear enough way that $cr(S|E)$ will be low. In other cases, the dispositions may be enough of a mix that it is hard to say.

Quineans might suggest that if $cr(S|E)$ is indeterminate in this way, and the subject later rejects S upon learning E, this should not count as a violation of conditionalization. If so, they then might suggest that for any S, there is some E such that $cr(S|E)$ is indeterminate in this way, and such that the subject could later reject S on learning E without violating conditionalization. Perhaps this sort of revisability is enough for their purposes?[34]

I do not think that this is enough, however. Cases of this sort seem to turn essentially on the subject's not being fully rational. If the subject is fully rational, then her dispositions to accept S on supposing E and on learning E should be the same, assuming no conceptual change. That is, if a fully rational subject rejects S on learning E and thinking things through, then if she were to have been presented initially with the supposition that E and had thought things through, she should have rejected S conditional on that supposition. To fail to meet this condition is a failure of full rationality, just as is an ordinary violation of conditionalization. So at best the Quinean has presented us with a kind of revisability that can be exploited only by subjects who are less than fully rational. Like the sort of revisability that can be exploited only by irrational subjects, this sort of revisability has no bearing on matters of apriority.

VIII. CONCLUSION

Quine is right that any statement can be held true come what may and that no statement is immune to revision. But as Grice and Strawson observe, these phenomena are quite compatible with a robust analytic/synthetic distinction and a robust notion of meaning. A Bayesian analysis reveals that Quine is not right that any statement can be held true come what may without conceptual change or irrationality, and likewise for revision. We can pin down the distinction between cases that involve conceptual change and cases that do not using either the method of intensions or Bayesian analysis.

---

[34] It is especially likely that ordinary subjects will lack credences $cr(S|D)$ involving the scenario specifications D discussed earlier, due to the enormous size of these specifications. This observation does not affect the use of conditional credences involving D to define intensions, as the credences used there are always idealized rational credences $cr'(S|D)$, for which the current issue does not arise. Where nonidealized credences are concerned, these cases will not yield cases of revisability along the lines in the text, because the subject will be incapable of learning that D.

The method of intensions characterizes intensions in terms of certain idealized dispositions, and uses this notion to distinguish between cases that involve conceptual change and cases that do not. In the central version that I examined, this method assumes the notion of apriority, so it does not ground that notion independently. Still, it shows how a framework involving apriority can accommodate all of Quine's data. And for the same reasons that most philosophers reject Quine's arguments in the first four sections of "Two Dogmas," no independent grounding is required.

The Bayesian analysis takes a step further and defends the a priori on partly independent grounds. This analysis assumes the notion of conditional probability and the normative notion of rationality to provide conditions for conceptual change, but it does not assume the notion of apriority. In effect, constitutive connections between rational inference and conceptual change are used to make inroads into the Quinean circle.

The conclusion should not be too strong. While I have responded to Quine's arguments against the a priori and the analytic, I have not provided a positive argument for the analytic/synthetic or the a priori/a posteriori distinction, and I have not tried to ground these notions in wholly independent terms.

One might be tempted to take things a step further still, and attempt to define apriority in terms of conditional probability and rationality. For example, one might suggest that a sentence S is a priori for a subject precisely when the ideal conditional probability $cr(S|D)$ is 1 (or: is high) for all scenario specifications D. But there will be residual issues. For a start, it is not clear that one can define the class of scenario specifications without using the notion of apriority. In addition, the thesis is subject to various potential counterexamples; for example, one might argue that when S is a mathematical truth and D specifies a scenario in which one is a poor mathematical reasoner, $cr(S|D)$ should be much less than 1. Much more would need to be said to handle these issues.[25]

[25] I address the second issue in *Constructing the World*, arguing that there is an idealization on which $cr(S|D)$ in this sort of case should be 1, where this idealization can be understood without appeal to the a priori. I think that the role of the a priori with respect to the first issue is more robust.

Still, we have seen that these notions can help us at least in diagnosing issues regarding meaning, conceptual change, and the a priori. And we have seen enough to suggest that Quine's arguments in the final section of "Two Dogmas of Empiricism" do not threaten the distinction between the analytic and the synthetic, or the distinction between the a priori and the a posteriori.

DAVID J. CHALMERS

Australian National University and
New York University

A MODEST PROPOSAL ABOUT CHANCE*

Before the year 1600 there was little uniformity in the conception of natural laws. In the seventeenth century, the rise of science, Newton's mathematization of physics, and the provision of strict, deterministic laws that applied equally to the heavens and to the terrestrial realm had a profound impact on the philosophical imagination. A philosophical conception of physical law built on the example of Newtonian mechanics quickly became entrenched. Between the seventeenth and twentieth centuries, there was a great deal of philosophical interest in probabilities, but probabilities were regarded as having to do with the management of opinion, not as having a fundamental role in science.[1] Probabilities made their first appearance in an evidently ineliminable way in the laws of a fundamental theory with the advent of quantum mechanics. Quantum probabilities have come to be called "chances" in the philosophical literature, and their interpretation has been one of the central problems in philosophy of science for almost a century. Hold-outs continue to insist that there must be an underlying, probability-free replacement for quantum mechanics, and Bohmians have had some success in formulating a deterministic alternative to quantum mechanics, but contemporary physicists accept that the probabilistic character of the quantum-mechanical laws is likely to be retained in any successor theory. While physics has adjusted itself comfortably to the existence of ineliminably probabilistic laws, philosophy has not managed to arrive at a stable interpretation of chance. The difficulty is that such an interpretation must satisfy a number of constraints. These constraints appear to be partially definitive of the concept, and it proves extraordinarily difficult to meet them simultaneously.[2]

* I would like to thank Huw Price, Guido Bacciagaluppi, and Alan Hájek for helpful comments. An audience at the Rutgers workshop on probabilities in statistical mechanics which included Shelly Goldstein, Nino Zanghí, Tim Maudlin, David Albert, and Barry Loewer provided an ideal sounding board. This paper is a descendent of a paper given at the Arizona Ontology Conference in 2009. Benj Helle's comments on that occasion were especially helpful. This work was made possible by a QEII fellowship from the Australian Research Council.

[1] Statistical mechanics did bring probabilities into physics, but not at the fundamental level and not in a way that challenged the strict necessity of the fundamental laws.

[2] "Chance" is sometimes used to refer to any form of objective probability. I use it narrowly to refer to quantum probabilities.

0022-362X/11/0808/416–442

Recently, there has been a push from a number of quarters in the philosophy of physics and the special sciences to recognize a form of probability in deterministic contexts. David Albert and Barry Loewer have argued for the objective nature of statistical-mechanical probabilities.[3] Elliott Sober has argued forcefully for the irreducibility of macro-probabilities and the need to recognize such probabilities.[4] I have argued that probabilities are indispensable quite generally in the formation of expectation, confirmation, and explanation, given any constraint on knowledge of the full microstate of the universe.[5] Wayne Myrvold and Hilary Greaves, reflecting on confirmation in an Everett universe, have found a role for probabilities in confirmation that is indifferent to whether or not the setting is deterministic. There are differences in motivation and detail among these defenders of objective, nonquantum probability.[6] All of them argue, however, that we have reason independent of quantum mechanics for recognizing a form of objective probability.[7]

The probability in question takes the form of a measure $\mathrm{Pr}_G(A/B)$, where A and B are (states or properties represented by) finite volumes of phase space and Pr_G is the probability that a random pick from (systems in) B will yield a system that is in A, or on an A-bound trajectory.[8] Pr_G defines both synchronic macroscopic probabilities (for example, the probability that an emerald is green or a woman over 40 has flat feet), as well as transition probabilities (for example, the probability that a baseball hitting a window at a speed of 40 miles per hour will break it).[9] In what follows, I will show how we can use

---

[3] Barry Loewer, "Chance and Determinism," in Yemima Ben-Menahem and Itamar Pitowsky, eds., "The Conceptual Foundations of Statistical Physics," *Studies in the History and Philosophy of Modern Physics*, xxxii, 4 (December 2001): 609–20; and David Albert, *Time and Chance* (Cambridge: Harvard, 2000). John Pollock, *Nomic Probability and the Foundations of Induction* (New York: Oxford, 1990) is an early precursor.

[4] Elliott Sober, "Evolutionary Theory and the Reality of Macro-Probabilities," in James H. Fetzer and Ellery Eells, eds., *The Place of Probability in Science* (New York: Springer, 2010), pp. 133–61

[5] Jenann Ismael, "Probability in Deterministic Physics," this JOURNAL, CVI, 2 (February 2009): 89–108.

[6] Or not-specifically-quantum probability. David Albert has argued that they ultimately may derive from quantum probability.

[7] To say that the probabilities are objective is to say at least that they are distinct from the subjective probabilities of the user of a theory. They are something that, to the extent that he is a good epistemic agent, *guides* his subjective probabilities.

[8] Ismael, *op. cit.*

[9] It can be derived from a distribution over initial conditions in deterministic contexts, but that is not by itself a reason for taking the distribution over initial conditions as the basic object.

As Albert points out, "once [a probability-distribution over possible microscopic initial conditions] is in place, all questions of what is and is not likely, all questions of what such probabilities to provide an interpretation of chance. One might wonder what is gained by defining chance in terms of another form of probability. The answer is that some very specific difficulties attach to the interpretation of chance, and defining chance in terms of Pr_G resolves these difficulties, while also explaining the role of chance in guiding belief and unifying the probabilities of quantum- and statistical-mechanical probabilities.

I. SINGLE-CASE AND GENERAL PROBABILITY

One of the most persistent sources of confusion in discussions of probability is the failure to distinguish general from single-case probability.[10] General probabilities apply to classes of events, and the basic form is conditional. The general probability of B among A's is written $\mathrm{Pr}(B/A)$.[11] Single-case probabilities, by contrast, apply to *particular* events rather than classes (that is, events that occur, or fail to occur, at a particular time and place), and the basic form is *un*conditional. Think of the difference between the general probability that an arbitrary roll of a pair of unbiased dice comes up double sixes and the single-case probability that *this particular roll* of *these particular dice* (which happen to be unbiased) on *this day* comes up double sixes. The former is conditional; it does not change over time; and there is an explicit reference class provided. The latter is unconditional; it has different values at different times; and there is no explicit reference class.

Single-case probabilities bear a direct connection to credence canonized in a principle that David Lewis dubbed the "Principal Principle."[12] General probabilities, by contrast, bear on credence only by way of their connection to single-case probabilities. The most important practical problem in assigning probabilities is that of determining which *general* probabilities to use to determine the single-case

---

was and was not to be expected, all questions of whether or not this or that particular collection of events happened merely 'at random' or 'for no particular reason' or 'as a matter of coincidence', are (in principle) settled," *op. cit.*, p. 10.

[10] General probabilities are also sometimes called "indefinite probabilities" and single-case probabilities called "definite probabilities."

[11] In logical terms, we can say that "Pr" is a variable-binding operator, binding the "x" in "$\mathrm{Pr}(Bx/Ax)$." I have suppressed the variables in the text.

[12] There has been dispute over the correct form of the principle, and various modifications have been proposed, one of which was accepted by Lewis before his death. I argued that the modifications are unmotivated in Ismael, "Raid! Dissolving the Big, Bad Bug," *Noûs*, xlii, 2 (June 2008): 292–307, but the disagreement need not detain us here. What matters for our purposes is something that is accepted by all parties, namely, that there is a tight link between chance and credence that any competent user of the concept knows at least tacitly

probability of a given event.[13] If I want to know the single-case probability that the next roll of a pair of dice will come up double sixes, for example, there are indefinitely many inequivalent general probabilities I might consult. There is the general probability that a roll comes up double sixes given that the dice are fair, or that a roll shows double sixes given that the immediately preceding roll was a double six, or that a roll shows double sixes given that I keep my fingers crossed.[14] Knowing which of these gives the right probability is highly nontrivial.

There are a number of views that one could have about the metaphysical relationship between single-case and general probability. One might hold that

(i) both are primitive forms of probability, neither reducible to the other,
(ii) general probabilities are definable in terms of single-case probabilities, or
(iii) single-case probabilities are definable in terms of general probabilities.

One of the difficulties in assessing accounts of probability is that some of them are offered as interpretations of single-case probabilities and others seem to be intended as interpretations of general probabilities, but they are compared as though they are competing accounts of the same object. This confusion runs deep in the literature, and there is little explicit clarity on the matter. Propensity interpretation and Humean reductions, for example, are seen most naturally as interpretations of single-case probabilities, whereas frequency interpretations are best understood as interpretations of general probability.

II. CHANCE[15]

By *chance*, I will mean the single-case probabilities that provide the link between the fundamental level of physical description in quantum mechanics and the measurement results that mark the points of empirical contact between theory and world. In (standard, nonrelativistic) quantum mechanics, states are represented by mathematical objects called "wave functions," and Born's Rule operates on wave functions to yield the chance that an observation or measurement on the system to which the wave function is assigned would yield a given result. Reiterated application of Born's rule generates a chance profile, that is, a probability assignment over results of possible measurements on the system. The chance of an event depends on the time at which it is assessed. Past events have a chance of 0 or 1; future events have chances in the interval between 0 and 1. More precisely, for any time t and any event e, we can assess the chance of e at t. If e is in t's past, it will have probability 0 or 1. If it is in t's future, it can take any real value in the interval between 0 and 1. An event that has a chance of 1 or 0 at t retains that value for all subsequent times. So, for example, if I hold a ticket to a lottery that will be decided by the result of a measurement carried out at noon on May 7, 2013, the chance of winning may have any value in the (closed) interval between 0 and 1 assessed at times leading up to that moment, whereupon it acquires a fixed value of 1 or 0.[16]

---

[13] Cognoscenti will recognize this as the Reference Class Problem.

[14] One very striking example of the lack of clarity about the distinction between general and single-case probabilities is that the term "Born Probability" is used without disambiguation to refer both to the general, conditional probabilities of the form Pr(α/Ψ) derived from Born's Rule and to the single-case, unconditional probability of α for a particular system in a state Ψ. Ask a physicist which one he means, and you are more likely than not to get a blank stare. It makes no difference when one is calculating, because the general, conditional probability of (α/Ψ) and the single-case, unconditional probability of some particular α in the future of a system in Ψ are quantitatively indistinguishable. But it makes a difference to interpretation The two have a different logical form; they bear different relations to categorical facts; they treat past events differently; and so on.

[15] Eric Winsberg and Roman Frigg use "chance" more widely than I, to refer to any form of objective physical probability. See Winsberg, "Laws and Statistical Mechanics," *Philosophy of Science*, LXXI, 5 (December 2004): 707–18, and Frigg and Carl Hoefer, "Determinism and Chance from a Humean Perspective," in Dennis Dieks et al., eds., *The Present Situation in the Philosophy of Science* (New York Springer, 2010), pp. 351–71.

These quantitative facts—that past events all have value 1 or 0, that an event that has value 1 or 0 at one time retains that value at later times—are important because they are not things that we had to discover about the distribution of chances at our world. They are things that anyone who "grasps the concept" of chance will be able to tell you, and they are our best clues to its nature. Another clue to the nature of chance is provided by the aforementioned link to belief that Lewis identified and formulated in his Principal Principle (PP).[17] PP says that if we know what the chance of an event e is and we have no magical sources of information from the future, we should adopt the chance of e as our credence. If chances are to guide belief, we have to have ways of forming beliefs about chances. That means that we have to have some idea of what counts as evidence for statements about chance, and that evidence will have to be connected to the truth conditions for those statements. What plays that role is a connection to frequency. Getting that connection right requires paying attention to the distinction between general and single-case probability. Recall that general probabilities apply to classes and do not generally have a time index. Single-case probabilities, by contrast, are unconditional, time dependent, and pertain to particular occurrences. Chance is a single-case, time-dependent probability. $\text{Ch}_t(e)$ is the chance that an event e has at t, where t is the time of assessment.[18] If e, the event that a coin toss to be made at noon on January 1, 2013 comes up heads, $\text{Ch}_{\text{now}}(e)$ is the present chance that the toss lands heads. In ordinary speech, we often suppress the temporal parameter, letting the conversational context decide the time at which the chance is assessed. If you ask me what chance I think you have of winning a bet predicated on the coin toss, for example, we both understand that you are asking for the chance that pertains at the time of asking. Now we are in a position to say how chance relates to frequency.

---

[16] This leaves open whether chancy quantum events ever attain extremal values. It asserts only the conditional that *if* an event e occurs at a time t, $\mathrm{ch}(e)$ after $t = 1$.

[17] Lewis regarded the Principal Principle as the sole constraint, believing that it told us "all we know about the concept of chance." There is controversy about the correct form of the principle. The differences will not matter for our purposes, but for proposals for revision, see Ned Hall, "Correcting the Guide to Objective Chance," *Mind*, n. s., CIII, 412 (October 1994): 505–17; Michael Thau, "Undermining and Admissibility," *ibid.*, pp. 491–503; John T. Roberts, "Undermining Undermined: Why Humean Supervenience Never Needed to Be Debugged (Even If It's a Necessary Truth)," *Philosophy of Science*, LXVIII, 3 (September 2001): S98–S108; and Peter B. M. Vranas, "Who's Afraid of Undermining? Why the Principal Principle Might Not Contradict Humean Supervenience," *Erkenntnis*, LVII, 2 (September 2002): 151–74 A defense of the original version of PP can be found in Ismael, *The Situated Self* (New York Oxford, 2007). The form adopted here is the one originally presented by Lewis, though he eventually adopted Hall's NP as a revision of the concept of chance. Lewis, "Humean Supervenience Debugged," *Mind*, n. s., CIII, 412 (October 1994): 473–90.

Bernoulli's law is a theorem that relates general probabilities to frequencies. It says that the relative frequency of A's in a typical ensemble of B's approaches $\Pr(A/B)$ as the size of the ensemble increases. This is good news; it gives us a necessary, probabilistic link between probabilities and frequencies. But there is a bad-news addendum: there are other theorems that tell us that that link cannot be strengthened. The possibility of divergence remains, no matter how large the ensemble.[19] This gives us the third constraint on the interpretation of chance, the chance-frequency link. The chance-frequency link says that if S is in state ψ, there is a necessary but approximate and probabilistic relation between the chance of observing e in an x-measurement on S and the relative frequency of e in x-measurements on systems in ψ.

We now have three clues to the nature of chance that we apply as constraints on interpretation;

(1) PP: one should set one's credence in a at t equal to the chance of a at t, no matter what else one knows, provided one has no magical information from the future.

(2) Quantitative constraints: the chance of an event after it occurs is always 1 or 0; an event that has value 0 or 1 at one time retains that value for all subsequent times.

CENTRAL LIBRARY
UNIVERSITY OF CALCUTTA

---

[18] Not to be confused with the time at which e occurs.

[19] Thanks to Alan Hájek for prompting this clarification.

(3) Chance-frequency link: the relative frequency of a's in a typical ensemble of systems in ψ approaches the general probability of a/ψ as the size of the ensemble increases, but the *possibility* of divergence in any finite ensemble remains, no matter how large the ensemble.[20]

(1) supplies an analytic connection to credence. (2) and (3) concern the relationship between chances and categorical facts. Categorical facts are the nonprobabilistic, nonmodal facts about what actually occurs. (2) places restrictions on combinations of chances and categorical facts. (3) provides a necessary but ineliminably approximate and probabilistic connection between chance and frequency.

III. THE INTERPRETIVE DILEMMA[21]

The constraints are easily satisfied individually, but conjointly they present a dilemma that destabilizes both the standard reductive and nonreductive accounts of chance. The problem is that they specify connections between facts about chance and facts about categorical events which seem, on the one hand, too loose to permit reduction and, on the other, too tight to let us treat them as distinct existences.

Why is the link between the chances and the categorical facts too loose to permit reduction? Bernoulli's law explicitly allows for the possibility that the chances may diverge arbitrarily far from the frequencies, which is a way of saying that it is a fact about the logic of chance that the very same distribution of actual events is *logically* compatible with an unlimited number of inequivalent chance distributions. The link between the actual pattern of events and the chances is irreducibly and irremediably probabilistic.[22]

---

[20] Support for all of these constraints, which we can treat as partially, provisionally definitive of chance, is that someone who denied them would not fully understand the concept. Someone who knew that George Bush won the election in 2004, for example, but denied that the present chance of his doing so is 1, or someone who knew what the chances were but did not think that they should guide his expectations, or someone who did not think that frequencies under the right conditions provided evidence for statements about chance, would not be regarded as understanding what chance is, or would be regarded as using the word in a different way.

[21] The interpretive dilemma distills a rather large body of literature to bare bones. For historical reasons, a good part of the philosophical discussion of chance is preoccupied with the compatibility of chance with Humean Supervenience (HS). HS is an independent metaphysical commitment, and it is not itself a constraint on the interpretation of chance. Lewis himself, the original defender of HS, held that the viability of HS hinged on the existence of an account of chance compatible with it, rather than the other way around.

[22] This raises questions about the Best-System Analysis of theories (BSA) discussed below, in connection with Loewer. The BSA purports to provide truth conditions for statements about probability, laws, and all of the apparently modal implications of a physical theory in terms of patterns in the manifold of categorical fact. It turns out, on this account, that to say that a statement L is a law is to say that the best overall

Why is the link between the chances and the categorical facts too tight to let us treat them as distinct existences? There are necessary quantitative constraints on the relationship between an event and the chance of its occurrence. Ask yourself whether *e* and the chances of *e* at different times can vary independently of one another. If these are distinct existences, there ought to be a possible world in which *e* occurs and the chance of *e* after the fact has any value we care to assign. So, for example, we ought to be able to conceive of a world in which a coin comes up heads and has a nonzero chance at later times of having come up tails. But the concept does not furnish a way of imagining the possibility. We can assign no intuitive content to the idea of such a world. This is a way of saying that certain combinations of categorical fact and values for chance do not make sense.[23]

This presents a dilemma for any account of the nature of chance. Chances have a peculiar, ontologically intermediate status that seems to frustrate both reduction to categorical facts and primitivism. From a formal point of view, chances look so much like ordinary physical quantities (they are represented by real-valued functions, they are assigned to particular systems, and they evolve in time) that it would be nice if we could treat chance as a primitive quantity, and say "so much the worse" for the Humean ban on necessary connections between distinct existences.[24] What is wrong with that reaction is that it fails to appreciate the regulative role that the Humean ban plays in these contexts. If we reject the Humean ban, we no longer have a way to recognize distinct existences. When we ask questions like, *what is chance?*, *what are colors?*, or *what is goodness?*, one of our goals is to provide a compact, nonredundant catalogue of fact. In this context, necessary connections of a non-nomological nature function as a sign of redundancy. Here is a familiar pattern of argument: Tom asserts that *A*-properties are metaphysically distinct from *B*-properties.

---

systematization of actual fact treats *L* as a law. And to say that the chance that a certain photon passes a filter is .99 is to say that the best overall systematization of actual fact entails that the chance that the photon passes the filter is .99. The BSA recognizes, however, that statements of law and probability have a fundamental status in physics. It is best understood as a meta-scientific view.

[23] Again, the claim here is that you should have trouble assigning intuitive content to the possibility of a world in which past occurrences have a nonzero present objective chance of not having occurred. It is not that you cannot dream up a scenario in which we might choose to use the concept in a way that allows for that possibility. It is that as we presently use it, that possibility has not been *provided for*. I take it for granted that this is as close as we come to analytic truth.

[24] Accounts that identify chances with propensities, explicitly denying that propensities are grounded in intrinsic categorical properties, seem to be doing that.

Alice responds that if *A*- and *B*-properties were distinct existences, there ought to be possible worlds in which the *B* facts are just as they are, but the *A* ones are wholly different, and *vice versa*. If there are no such possible worlds, Tom is wrong. The Humean ban on necessary connections between distinct existences functions in this pattern of argument as a test for ontological redundancy, and without it we no longer have a methodological foothold for addressing claims of metaphysical distinctness.

A rather large philosophical literature has built up around the interpretation of chance. Lewis's papers in the early 1980s brought the topic under special scrutiny because of the difficulty he saw with incorporating chance into his ontology.[25] There is a small collection of well worked-out interpretations. There are frequentist accounts, propensity accounts, primitivist accounts, nomic accounts, Humean accounts, and subjectivist accounts. Each of these, however, suffers from well-known problems. Each comes with one or another bullet to bite. Those working for reductions are responding to the fact that *e* and the chance of *e* do not behave like distinct existences. Primitivists are responding to the fact that reduction seems blocked by the ineliminably probabilistic character of the link between chances and frequencies. But neither can respond appropriately to pressures from the other side. As things stand, by general consensus, there is no entirely satisfactory interpretation of chance.[26]

IV. THE PROPOSED SOLUTION

There is a lesson that one learns in physics or math when one feels backed into a corner with a difficult problem: try the obvious solution. I want to suggest that there is a way out of these difficulties that is so obvious, once we have independent grounds for recognizing the existence of general probabilities, it can strike one as almost trivial. It takes a little unpacking to see exactly what is accomplished. The conceptual character of the quantitative constraints that an event's occurrence places on subsequent chances provided clear symptoms

[25] See Lewis, *op. cit.*

[26] There are many excellent critical discussions and surveys of the standard influential interpretations of chance. See Pollock, *op. cit.*, Michael Strevens, "Probability and Chance," in Donald M. Borchert, ed., *Encyclopedia of Philosophy*, 2nd ed. (Detroit: Thomson Gale/Macmillan Reference, 2006); and Craig Callendar, "The Emergence and Interpretation of Probability in Bohmian Mechanics," in Frigg and Stephan Hartmann, eds., "Probabilities in Quantum Mechanics," *Studies in the History and Philosophy of Modern Physics*, xxxviii, 2 (June 2007): 351–70. Alan Hájek provides an especially nice general survey in "Interpretations of Probability," *The Stanford Encyclopedia of Philosophy*, ed. Edward Zalta (Spring 2010). URL: http://plato.stanford.edu/entries/probability-interpret/.

of nonbasicness, so reduction is needed, but not reduction to categorical facts. We reject both primitivism and reduction to categorical facts and define the chance of *e* assessed at *t*, written $\mathrm{Ch}_t(e)$, as follows:

Def: $\mathrm{Ch}_t(e) = \mathrm{Pr}_G(e/\text{pre-}t \text{ history})$

where $\mathrm{Pr}_G(A/B)$ is the general probability that a random pick from the *B*'s will yield an *A*.[27]

It is easy to see that Def satisfies the second and third of the constraints on the interpretation of chance. It follows immediately from the definition that the chance of any event assessed after its occurrence will always be 0 or 1. And since Pr_G is a probability function, it is going to satisfy the probability axioms, and the chance-frequency link will hold. The nontrivial part of establishing that Def satisfies the constraints on interpretations is saying why the measure defined by Def should play the role characterized by PP in guiding belief. Recall that PP said that provided we have no magical information from the future, the degree of belief we assign to *e* should equal the chance of *e*. But there are any number of functions of Pr_G, any of which have the right form to play that role. Consider, for example,

$\mathrm{Pance}(e) = \mathrm{Pr}_G(e/\text{pre-2001 history of Australia})$
$\mathrm{Fance}_t(e) = \mathrm{Pr}_G(e/\text{post-}t \text{ history})$
$\mathrm{Trance}(e) = \mathrm{Pr}_G(e/\text{all of history})$

In contrast to these functions, what makes Def peculiarly suited for the role carved out by PP? The answer is that chance is the only one of these functions that trumps *all* and *only* historical information. For creatures like us, historical information is in principle available, whereas information from the future is out of bounds. Chance guides belief because it is probability conditioned on all information that is in principle available to a situated agent. As such, it is the only function of Pr_G that trumps all and only such information. Belief guided by Pance(*e*) and Fance(*e*) would lead us to assign probabilities other than 1 to historical events that we know occurred, or nonzero probabilities to historical events we know did not occur. Belief guided by Trance(*e*) will not lead to those kinds of mistakes, but there is a different problem. Information about the "trances" of future events is in general beyond our ken and hence unavailable to guide present belief.

---

[27] The definition assumes that the general-probability distribution ranges over all possible (historical) events, so that $P(E|H)$ is well defined for any possible world history *H*. So, strictly speaking, it applies in the first instance to theories that treat the world as a closed system, and hence to cosmological theories. It is in that context—a deterministic world—that it seemed probabilities were eliminable. In the case of open systems, probability appears explicitly in the form of a distribution over exogenous variables, and the status of that distribution depends on the context. In some cases, it is explicitly epistemic. In others, it is statistical.

V. $\mathrm{PR}_G(A/B)$

Def does not provide a reduction of facts about chance to nonprobabilistic facts. It provides an interpretation of chance in terms of a form of general probability. The burden of argument then falls on questions about Pr_G. What is it? What reason do we have to believe in it? What sort of progress is made by reducing chance to $\mathrm{Pr}_G(A/B)$? Addressing these in turn: Pr_G defines the notion of randomness according to a particular theory. $\mathrm{Pr}_G(A/B)$ is the general, conditional probability that a random pick from some proscribed volume of phase space (B) will yield a system whose state—at that time or some later time—falls in another such volume (A).[28] Grounds for the recognition of Pr_G come from a number of sources. Loewer and Albert have argued for recognition of a form of objective probability to ground statistical-mechanical probabilities.[29] Sober has shown that evolutionary theory makes use of objective macro-probabilities and has given general arguments for the recognition of objective probabilities.[30] We do not need the examples of statistical mechanics, evolutionary theory, or special sciences to make the case. There can be little question that some measure is tacitly and routinely invoked as a matter of practical necessity in classical contexts with deterministic laws, because without such a distribution the laws are virtually impotent to govern expectations or generate testable predictions.[31] Global deterministic dynamical laws determine what is *possible*, but they do not tell us how to divide opinion among the possibilities. In any realistic situation, there are indefinitely many present states for the universe as a whole that are epistemically possible relative to our knowledge, and hence indefinitely many future states into which the universe can evolve.[32] The same is true for any proscribed subsystem of the universe because of the impossibility of perfectly isolating it from its environment. Any proscribed subsystem is in principle subject to innumerable possibilities of unanticipated interference. We can lower the probability of interference, but the possibilities are ineradicable. Laws that tell us only which future states are possible given our present knowledge will give an infinite, undifferentiated set of possibilities with no distinction between significant and negligible possibilities or between relevant and irrelevant influences.[33]

---

[28] I remain agnostic on whether the backward probability $\mathrm{Pr}_G(A/B)$ where B occurs after A is essential. In an indeterministic setting, the backward probabilities may not be given. In a deterministic setting, they follow from the forward probabilities. See Guido Bacciagaluppi, "Probability, Arrow of Time and Decoherence," in Frigg and Hartmann, eds., *op. cit.*, pp. 439–56.

[29] See Loewer, *op. cit.*; and Albert, *op. cit.*

[30] See Sober, *op. cit.*

[31] This is a qualified conclusion. One can imagine laws that retain their predictive power even in the absence of precise knowledge, but the Newtonian dynamical laws do not. So the claim is that global determinism *by itself* places very weak constraints on the evolution of local subsystems of the universe. For examples of laws that do retain predictive power even in the absence of precise knowledge, consider laws that entail that any system takes the shortest path through phase space at a fixed speed to a final state S, no matter what its starting state and in a manner entirely immune to influence. To multiply examples, one just has to limit interaction and influence.

In practice, we deal with ignorance by randomizing over the values of unknown variables. Talk of randomness is not well defined without a choice of Pr_G. As we saw, Pr_G effectively *defines* what is meant by "random" and (hence) guides the assignment of probabilities where there is ignorance.[34] To see this, let W be a collection of physically possible worlds (or physically possible histories for the world), and let W^* be the set of worlds in which the system of interest s exists and possesses all of its known properties. If W^* contains more than one world, there will be indefinitely many probability distributions over W^*. When we talk about randomizing over the values of unknown variables, we are talking about choosing a distribution over W^*. It does not help to adopt an indifference principle; indifference principles can be applied consistently only over *equiprobable* alternatives. And it does not help to say that we should give a flat distribution, because there are different ways of partitioning any set of possibilities, and a flat distribution over one partition will yield a nonflat distribution over another. We could demand a flat distribution over the most fine-grained partition. This may deliver a univocal answer in the finite case, but it will not do so in the denumerably infinite or continuous cases. There is a long history of attempts to justify the choice of $\mathrm{Pr}_G(A/B)$, and different views about whether and how the choice can be justified, but there is no way around the fact that some choice is needed. The point is prior to and independent of any particular account of the nature of chance.[35]

---

[32] The reference is to: "Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes." Pierre-Simon Laplace, *A Philosophical Essay on Probabilities*, trans. Frederick William Truscott and Frederick Lincoln Emory (New York: Cosimo, 2007), p. 4.

[33] One can, of course, deny that a deterministic theory generates expectation, holding that it tells us only what is possible and impossible. Everything else is done by the subjective probability distribution which one brings to a theory. In practice, that is not what we do. A canonical distribution is assumed, usually tacitly, that is not regarded as entirely discretionary. The explicit content of a theory needs to include a probabilistic postulate to lend the theory some testable content.

[34] A little more precisely, let S be a system of interest, let C be a summary of everything that is known about S, and let $A = \{A_1, A_2, A_3, \ldots\}$ be an unknown variable with values $\{A_1, A_2, A_3, \ldots\}$. Randomizing over A-values, holding C fixed, involves assigning each value of A the probability that a random pick from systems that satisfy C will yield that value. But that is just to say that it involves assigning the probability distribution over $\{A_1, A_2, A_3, \ldots\}$ generated by $\mathrm{Pr}_G(A_i/C)$.

Choosing Pr_G is equivalent to choosing a standard that identifies partitions of W^* that divide it into equiprobable classes. In physical contexts, this standard gets embodied in the metrical structure of the phase space, and such strong intuitive prejudices operate in dividing possibilities into equiprobable classes that the choice of $\mathrm{Pr}_G(A/B)$ often goes unmentioned and unnoticed. In classical mechanics, it did not come to light until disputes in the foundations of statistical mechanics bought it under scrutiny. Nowadays, as a result of those disputes, there is recognition among philosophers of physics that a form of statistical probability is both compatible with determinism and needed to derive a dynamics for larger than point-sized volumes of phase space. We can picture these higher-level dynamics as migration patterns which tell us how a class of systems distributed in a given way across a finite volume of phase space typically redistribute themselves over time.[36] To generalize the conclusion that the higher-level migration patterns are needed to derive expectations not only in the specialized context of statistical mechanics but in virtually any real-life calculation, we need only observe that the scientist working with any real-world system is always working with a system whose state can only be isolated, as an epistemic matter, in a finite volume of the universal phase space.[37] He appeals to the general information embodied in the large-scale migration patterns to make up for his lack of specific information.

What lends $\mathrm{Pr}_G(A/B)$ empirical content is the link to frequencies given by the Bernoulli principle. The bad-news addendum to the principle blocks a direct reduction to frequencies, but the probabilistic link is enough to make Pr_G empirically meaningful. I argued in *The Situated Self* that Pr_G should be treated alongside the physical laws as a primitive postulate whose correctness is justified by the success of the theoretical package of which it is a part. The success of the package as a whole confers whatever confirmation it possesses upon its elements. And confirmation is judged not by compatibility with the laws, but by the ability of the theory to predict the evidence with high probability. Note, however, that to say that $\Pr_G$ is a primitive theoretical postulate is to remain neutral on whether it can be reduced, eliminated, or analyzed as a particular form of idealized subjective probability at a meta-scientific level. It is possible for reasonable and well-educated philosophers to accept that laws of nature are physically fundamental but disagree about what laws are, whether facts about laws can be reduced to or analyzed in terms of some more fundamental set of facts, or whether we have good reason to believe that there are any laws. In the same way, it is possible for reasonable and well-educated philosophers to agree that a probability measure of the form $\Pr_G(A/B)$ is physically fundamental and disagree about what it is, whether facts about it can be reduced to or analyzed in terms of nonprobabilistic facts, and even whether we have good reason to believe in its existence.[38] Questions about which structures are *physically fundamental* aim to identify the components needed for a working theoretical package. They can be treated independently of questions about the metaphysical nature or status of those components.[39]

---

[35] This is a very quick summary of arguments from Ismael, "Probability in Deterministic Physics."

[36] To derive these higher-level migration patterns for any distribution, it suffices to say how a flat distribution evolves.

[37] Although it is still counter to orthodoxy, the idea that objective probability may be more than just a quantum-mechanical phenomenon is gaining some currency.

But if $\Pr_G$ is accepted as a primitive postulate, the question of what progress is made by reducing chance to $\Pr_G(A/B)$ becomes pressing. The answer is that the definition of chance in terms of $\Pr_G(A/B)$ reduces it to a form of statistical probability that also underwrites the probabilities of statistical mechanics, explains the epistemic role of chance, and, since $\Pr_G(A/B)$ bears no necessary connection to categorical facts, resolves the dilemma generated by treating chance as fundamental. Once we clear up some hard-to-pinpoint confusions on the interpretation of chance, which stem from a failure to distinguish general from single-case probability, and appreciate the need for general probability in deterministic and indeterministic contexts, Def tells us that chance is nothing more than a form of single-case probability derived from Pr_{G} by conditionalization on the past. We can then focus our attention on the interpretation of Pr_{G}. By defining chance in terms of Pr_{G}, Def preserves the temporal symmetry of the fundamental theoretical postulates. It is also a nice feature of Def that Pr_{G} is not time dependent. There is a very real worry that because chances of past events always have values of 0 or 1, while chances of future events range in the (closed) interval between 0 and 1, admitting chance into physical theory recognizes a fundamental form of temporal asymmetry; it says that the past is in some fundamental respect different from the future.[40] It should be noted, however, that although Def reduces chance to a form of general probability, it is not the sort of reduction that Einstein envisaged. Einstein's hope was to reproduce the quantum-mechanical probabilities from deterministic laws and a probabilistic postulate on the model of statistical mechanics. The proposal here accepts the ineliminably probabilistic character of the dynamical laws but spreads the guilt by arguing that determinism offers no respite from probability. Even deterministic theories cannot make do in practical terms without a specification of Pr_{G}.

---

[38] This last depends on whether x's being a fundamental postulate of science provides a reason to believe that x exists. Not all philosophers think it does.

[39] The recent debate between Maudlin and Loewer about the nature of laws is instructive in this regard. See Tim Maudlin, "Why Be Humean?" and "A Modest Proposal Concerning Laws, Counterfactuals, and Explanations," in *The Metaphysics within Physics* (New York: Oxford, 2010); and Loewer, *op. cit.* Both agree that laws are physically fundamental structures, as do, for example, David Armstrong, *What Is a Law of Nature?* (New York: Cambridge, 1983); and Bas van Fraassen, *The Scientific Image* (New York: Oxford, 1980). But they hold a wide range of views about the nature of laws. Maudlin denies that there is any physics-independent platform from which to raise questions about ontology, and regards physically fundamental structures as ontological bedrock. Loewer follows Lewis in thinking that laws (and everything else that exists) reduce to patterns in the Humean manifold of events. Armstrong holds that laws are relations between universals. One might just as well accept that they are primitive physical postulates but hold that they are convenient fictions, nodes in an uninterpreted calculus for deriving predictions about observable structures, or the real structures that underlie and explain the world's observable properties

Why, and in what sense, is Pr_{G} objective? It is objective at least in the sense that it is not discretionary. I have said that it is a form of objective probability. There is a lack of clarity in the literature about what the claim of objectivity amounts to. For me, it means that it is to be conceived as part of the content of a theory, defined independently of the subjective probabilities of its users.[41] Pr_{G} is implicated in prediction, confirmation, and explanation. These ought to flow from the theory rather than the judgment of the user. And the single-case probabilities derived from Pr_{G} under conditions of partial information *guide* rather than *describe* subjective probabilities. Winsberg and Frigg have objected recently to what they regard as ontologically inflated interpretations of statistical-mechanical probabilities. They argue that the only stable position construes them as objectively constrained epistemic probabilities suited for agents with our particular epistemic profile. Pr_G is objective, but not ontologically inflated in their sense. I agree with their arguments, and chance (on my interpretation) is well suited to their view as a form of statistical probability poised to guide the opinions of agents with our particular mix of ignorance and knowledge.

---

[40] This is a controversial topic Both the characterization of temporal symmetry and the question of whether the existence of chance violates it are unsettled. See Bacciagaluppi, *op. cit.*

[41] This leaves a lot open. It allows, but does not require, an interpretation of theories as embodiments of expert opinion, so that chances turn out to be credences of an expert believer.

VI. MAUDLIN AND LOEWER

This is a good point to relate what I have said to some recent work of Tim Maudlin and to say a little more in detail about Loewer. Maudlin's view is a form of what can be called "nomic probability."[42] In nomic accounts, the stochastic laws define a form of general, conditional probability, and the general probability is used to define chance by conditionalization on history. Maudlin writes:

> Let us take a deterministic FLOTE [fundamental law of temporal evolution] and adjunct principles that operate in a special relativistic space-time. Take a surface that cuts every maximal timelike trajectory in the space-time exactly once (a Cauchy surface)....Boundary values can be specified on this surface, such as the distribution of particles, intensities of fields, etc.....If the FLOTE is deterministic in both the past and future directions, then the boundary values will determine a unique distribution of the physical magnitudes through all time. Such a distribution describes a physically possible world relative to those laws.[43]

When we have deterministic laws and a specification of boundary values rich enough to define a Cauchy surface, the laws constrain history in the future direction to a single model and no probabilities are needed. But, he continues:

> If the FLOTE is stochastic....specific boundary values on the Cauchy surface yield not a single model but a set of models, corresponding to all of the outcomes permitted by the laws. Furthermore, the set of models consistent with some boundary values is invested with a metric over measurable subsets, a measure of how likely it is, given those boundary values, that the world will evolve into one of the members of that subset.[44]

The case that Maudlin does not discuss, and the one that brings the need for the fully general Pr_G into relief, is the deterministic case in which we do not have enough information to identify a unique Cauchy surface. If epistemic probabilities are to be empirically constrained, that is, if it is not an entirely discretionary matter what probabilities are assigned when we are dividing opinion among multiple epistemically possible models, we need an objective metric over measurable subsets of physically possible trajectories. This is so whether or not the fundamental laws of evolution are deterministic. In these cases, we need higher-level laws that describe what I referred to earlier as migration patterns over phase space, that is, laws that tell us not how systems in specific microstates are bound to evolve, but how systems whose states fall within some volume typically distribute themselves across that volume and redistribute themselves over time.[45] So, I agree with proponents of nomic probability that the notion of a stochastic law should be treated as more primitive than that of chance, and that chance (defined as "objective, single-case propensity"), should be defined as Maudlin says, effectively by conditionalizing on the past. But I disagree that probabilities are eliminable when we have deterministic laws. Nontrivial probabilities in deterministic contexts invoke the same kind of measure over physically possible trajectories that Maudlin recognizes in the indeterministic case, and they are implicated in every real-world derivation of expectation from theory.[46]

---

[42] John Pollock coined the term. He was an early proponent of nomic probability.

[43] Maudlin, "A Modest Proposal Concerning Laws, Counterfactuals, and Explanations," p. 18. One can see the nod to Maudlin's paper in my title. Maudlin includes a footnote acknowledging qualifications about constraints and boundary conditions, but he leaves it at that.

[44] *Ibid.*, pp. 18–19

Loewer states his position in terms of a philosophical framework that I have reservations about, a development of Lewis's Best-System Analysis (BSA), which aims to reduce facts about laws, chances, and physical modality quite generally to facts about patterns in the manifold of categorical events.[47] But if we put aside the claim that the BSA provides truth rather than, for example, *acceptance* conditions for a physical theory, our views are almost indistinguishable. There are some differences of detail, but we agree that probability is compatible with determinism, and that a probabilistic postulate included alongside the laws as a component of a theoretical package can be used to define chance.[48]

---

[45] This does not depend on epistemic probabilities that are flat over the initial volume of phase space. Even if your epistemic state is represented by a distribution that is heavily clustered in one corner of the volume, for example, you need to know how ensembles clustered that way typically evolve.

[46] It is worth noting that Maudlin's definition of chance applies smoothly to the more general context...

[47] See Loewer, *op. cit.* My reservations stem from the fact that to be successful, the BSA has to recognize that even though the acceptance conditions of a theory can be given in terms that advert only to the pattern of actual events, statements about law and probability have modal implications that cannot be reduced to claims about actual events. In my view, the BSA has the wrong account of modality.

[48] I would say something stronger, namely, that a probabilistic postulate is not just compatible with but practically indispensible for determinism, and that there is no measure of fitness with evidence without a probabilistic postulate.

VII. WHY TREAT PR_G AS THE BASIC OBJECT?

I have said that what we need in order to derive expectations in practice—outside the context of pen-and-paper problems in which all of the information is artificially stipulated—are migration patterns over phase space which deliver the probability that a system which starts out in some finite volume ends up in another at some later time. These are not derivable in any context, deterministic or otherwise, without a probability measure put in somewhere. But one might ask: why treat $\mathrm{Pr}_G(A/B)$ as the basic object? Why not some other form of probability? Why not, for example, a probability distribution over initial conditions?

The first reason is that the idea of a probability distribution over initial conditions is notoriously problematic. What sense—empirical or conceptual—does it make to talk about a distribution over initial conditions if the universe occurs only once? $\mathrm{Pr}_G(A/B)$, by contrast, gets empirical content from a probabilistic link to frequencies in finite actual ensembles, for example, the frequency of tosses of American quarters from 1976 that land heads, the frequency of vegetarians among French people, the frequency of smokers under 50 that get cancer.[49] These finite ensembles are the primary site of empirical application of probabilistic concepts. A distribution over initial conditions gets whatever empirical contents it has indirectly, from the constraints it places on them. Tracing all probability to a distribution over initial conditions of the universe obscures its empirical content, effectively reducing it to an empirically empty case.

The second reason for treating Pr_G as basic is that it cannot be generated from a distribution over initial conditions in indeterministic contexts. When dynamical evolution multiplies the number of possibilities, an initial distribution will not determine a final one. So it is a special feature of deterministic laws that they can generate a distribution at later times from a distribution over initial conditions. This is a good reason for thinking that Pr_G is the more general and basic form.[50]

VIII. NOVELTY

If the novelty introduced by indeterministic laws does not consist in the appearance of a new form of probability or even a new role for probabilities in inductive inference, wherein does it lie? It consists in the elimination of a degenerate case present in a deterministic setting. I have argued that probabilistic assumptions play an ineliminable role whenever we are working with volumes of phase space from which multiple physically possible trajectories emerge. In deterministic theories, the number of such trajectories goes to 1 as the size of the volume goes to 0. So eliminating historical ignorance reduces all prediction to the special case in which there is only one possibility. In indeterministic theories there are multiple trajectories through every point, so there is no degenerate case, and even if we eliminate historical ignorance, a measure is needed to set credence. There is a way of putting this that makes contact with the technical work on hidden variables in quantum mechanics. In deterministic theories, even when we are working with imprecise or incomplete information, there is a more fine-grained description that locates the system in an underlying space with a single trajectory through every point. In the indeterministic case, there is no such underlying space, no fine-grained description with a single trajectory through every point that will let us reconstruct the probabilities at the higher level from a distribution over variables whose values are (in principle, jointly) measurable. We have to formulate the laws using a space that has multiple trajectories through some points, and that means that if we are to end up with an objective measure over possibilities, probabilities have to appear explicitly in the statement of the laws.

---

[49] See Detlef Dürr et al.,"Bohmian Mechanics," in Borchert, ed., *op. cit.*; Dürr, Sheldon Goldstein, and Nino Zanghí, "Quantum Equilibrium and the Origin of Absolute Uncertainty," *Journal of Statistical Physics*, LXVII, 5/6 (1992): 843–907; and Callendar, *op. cit.*

[50] It is a special feature of bi-deterministic laws that they can generate a distribution at one time from a distribution at any other. The Newtonian laws are bi-deterministic

Can't we reinstate determinism trivially by "discerning" variables so hidden that they cannot be measured and use them to parameterize phase space with an intrinsic measure that generates the Born measure over quantum-mechanical state descriptions? This is a natural suggestion, and the one that Einstein explored until his death. Under such an arrangement, quantum-mechanical state descriptions would correspond to finite volumes of the underlying space in the same way that thermodynamic state descriptions correspond to finite volumes of the phase space of statistical mechanics. The problems with the suggestion are the topic of a vast and technically formidable literature exploring the possibilities for hidden-variable interpretations. There are unsettled questions, but we do know that any such interpretation will exhibit nonlocality or contextuality of some form at the fundamental level, and it is very difficult both in physical terms and conceptually to work out the resulting implications.[51]

---

[51] This is so at least so long as space-time is retained as the fundamental arena of physical description. There are some interesting suggestions from multiple quarters that treat space-time as emergent and reinstate determinism and locality as the fundamental. For an excellent discussion of hidden-variable results, see Carsten Held, "The Kochen-Specker Theorem," *The Stanford Encyclopedia of Philosophy*, ed Edward Zalta (Winter 2008) URL http://stanford.library.usyd.edu.au/entries/kochen-specker/.

All of this raises the question of why the interpretation of chance has operated under the dogma that chance is physically fundamental. It is not just that probabilistic assumptions work behind the scenes in deterministic contexts. Even where probabilities are in explicit use there is confusion about their status. I believe it is easy to slip into thinking that the probabilities represent facts about our epistemic states in part because they are only needed in the deterministic case when there is historical ignorance. But that is not right. What they represent are objective, statistical facts about migration patterns of typical ensembles in the general population, and we rely on these general facts when our specific knowledge gives out. Another reason, however, is a holdover from a conception of physical law built on the Newtonian example, according to which dynamical evolution is a species of causal *production*, so that dynamical laws tell us how one state of the universe produces the next in a manner that is temporally asymmetric and involves ideas of causal sufficiency.[52] In a relativistic setting, this conception of laws does not fit. Laws are reinterpreted as global constraints with no intrinsic temporal direction and no special connection with time. The distinction between law-like and contingent features of the world is central to any physical theory, but from the perspective of this new conception of law, it is an inessential and quite special feature of the deterministic case that contingencies can be isolated on a space-like hyperplane, so that once we fix the state of any such hyperplane the laws determine everything else. From this new perspective, there is no reason to expect the distinction between law-like and contingent features of the manifold to line up in this simple way with the division of space and time. In an indeterministic setting, contingency is distributed throughout space and time. From a relativistic point of view, this seems a more natural arrangement. It would seem a surprising accident if the world *did* turn out to be deterministic.

IX. METHODOLOGICAL REORIENTATION

I have been arguing that the philosophical literature has treated chance as a surprising and anomalous protrusion of probability into an otherwise probability-free environment, whereas in fact it is a special case of something more fundamental and perfectly pervasive, the tip of a probabilistic iceberg that is present but remains largely below the surface in deterministic settings. Accepting this prompts a methodological reorientation, shifting philosophical attention from chance to Pr_G. This shift of attention has implications for how we think of physical probability. First, unlike chance, Pr_G is not inherently dynamical. If we can talk about the probability of evolving from one subspace of phase space into another, we can talk equally about the degree of overlap between two subspaces (that is, the probability of being in *B*, given that one is in *A*). Second, Pr_G has no intrinsic temporal asymmetry. It defines nontrivial transition probabilities in both directions, from past to future and future to past. Third, we no longer have to suppose that there are two distinct types of physical probability: chance and the micro-canonical probability measure invoked by statistical mechanics. We have a single measure, Pr_G, from which both can be derived.[63] Finally, rejecting the dogma that chance is physically fundamental resolves the interpretive dilemma that chance suffers by focusing interpretive attention on a form of general statistical probability rather than single-case, dynamical probabilities. Statistical probability provides a much more intuitive entry point into the circle of probabilistic concepts.

---

[52] Maudlin, "A Modest Proposal Concerning Laws, Counterfactuals, and Explanations," does a very nice job of articulating this conception, which he himself endorses.

X. APPLICATIONS

I have spoken nonrelativistically for intuitive ease, but I want to pause to say how to translate into a relativistic setting and mention a couple of applications. To translate into a relativistic setting, we identify systems with world lines, define chances at points, and substitute "contents of past light cone" for "history." So now we have:

Def*: $\mathrm{Ch}_p(e) =_{\mathrm{def}} \mathrm{Pr}_G(e/\text{the contents of } p\text{'s past light cone})$

Traditional Laplacian definitions characterize chancy events as those that are undetermined by the dynamical laws from the preceding state of the universe. This leaves no room for chancy events in a universe governed by global deterministic laws, like the one proposed in various Everettian interpretations of quantum mechanics. One of the virtues of Def* is that it separates the existence of chance from determinism, tying it to the light-cone structure. If we replace the Laplacian definition with:

> The occurrence or nonoccurrence of an event is *chancy* just in case it cannot be predicted with certainty by application of the dynamical laws to the contents of (that is, to the open set of events that lie in) its past light cone,[64]

---

[63] On Albert's formulation, classical statistical mechanics consists of three postulates: (1) the fundamental dynamical laws, (2) a uniform probability distribution—the micro-canonical distribution—over the possible phase points at the origin of the universe, and (3) a statement characterizing the origin of the universe as a special low-entropy condition. $\mathrm{Pr}_G(A/B)$ is equivalent to the micro-canonical probability distribution. And if we put a low-entropy initial state in for *B*, it will give us the thermodynamic laws. See Albert, *op. cit.*

[64] For ease of definitions, I include *e* in its future light cone but not its past.

we preserve a connection between chance and predictability but break the connection with determinism, because the connection between predictability and determinism is lost in the context of a light-cone structure that imposes greater restrictions on the availability of information. The result is that if we apply the definition in an Everett universe, we get chancy events despite the determinism of the global dynamical laws. To see how this works, suppose that I am an observer in an Everettian universe about to carry out a spin measurement on a spin-half particle in a single state. Suppose that there is actual splitting into separate, spatiotemporally disconnected branches of the universe and that there is no evolution between the time of the measurement and observation of a result. There will be two actual observations by downstream descendents of myself. One (let us call her "Jenann-up") will observe spin-up. The other ("Jenann-down") observes spin-down. Both observations will be chancy according to the definition, because neither Jenann-up nor Jenann-down will be able to derive the content of her particular observation with certainty from nomic probabilities by conditionalizing on the (pre-observation) contents of her past light cone. The best that she will be able to do is assign probabilities in accord with the typicality of those results in physically possible trajectories that match the past light cones of the observation event. And we can give quantitative significance to the Born probabilities. They function just like the general probabilities in a classical setting, as derived from a metric over measurable subsets of physically possible trajectories obtained by conditionalizing on the contents of the back light cone.[55] The probability of observing α in an O-measurement on a system in state Ψ is $\Pr(\alpha/O\&\Psi)$.

Another nice application of specialized interest is that $\Pr_G$ gives us a natural interpretation of the notion of typicality that plays a central role in the Dürr, Goldstein, and Zanghí (DGZ) version of Bohmian Mechanics.[56] DGZ explicitly disavow a construal of typicality in terms of probability on the grounds that it is empirically empty to talk about the probability of initial conditions. As Craig Callendar writes:

> An instinctive negative reaction to assigning probabilities to initial conditions is natural and probably even healthy. DGZ are understandably reluctant to dub the universe probable, for it invites quasi-theological pictures of supernatural beings picking the universe out of an urn.[57]

---

[55] In this case, all physically possible trajectories compatible with a pre-measurement history are actual, but they work just as in Maudlin's account of the indeterministic case

[56] Dürr, Goldstein, and Zanghí, "Quantum Equlibrium and the Origin of Absolute Uncertainty."

[57] Callendar, *op. cit.*, p. 367.

As mentioned above in connection with reasons for taking Pr_G rather than a distribution over initial conditions as basic, Pr_G gets its empirical content from the constraints it places on observed statistics in smaller than universe-sized samples.

XI. ADDRESSING OBJECTIONS

A couple of objections are worth addressing before closing. The first objection is that Def tells us that the chance of *e* at *p* is the single-case probability obtained from Pr_G by conditionalizing on historical information accessible from *p*. In fact, nobody ever possesses full historical information, so chances are epistemically inaccessible in practice. So how can they guide belief in the manner dictated by PP? The answer is one that I have defended in more detail elsewhere.[38] I hold that PP correctly captures the conceptual connection between chance and belief, but in practice what applies is a generalization of PP that accounts for historically based ignorance by forming a mixture of chance functions obtained by conditionalizing on epistemically possible histories whose weights reflect our credence that those histories are actual. In short, when we do not know what the chances are, belief is guided by our best estimate of the chances.

An interesting addendum to this response arises from a reply by Alan Hájek. Hájek observed that the response I just gave threatens to undermine my answer to the earlier question of why $\mathrm{Ch}_t(e)$ is peculiarly suited to play the role carved out by PP in guiding belief. Consider in particular the "trances" which are obtained from Pr_G by conditionalizing on all of history: $\mathrm{Trance}(e) = \mathrm{Pr}_G(e/\text{all of history})$. The reason for denying that these guide belief in the manner dictated by PP was that if we do not have crystal balls, we do not in general know what the trances are. But once we allow that in practice we do not know what the chances are either, the door is opened for arguing that the trances do as good a job guiding belief as the chances.[39] After all, Trance(*e*) agrees with $\mathrm{Ch}_t(e)$ on past events, and it does at least as well as the chances on future events. And it is certainly true that *if* we know the trances we ought to adopt them as our credences, no matter what other information we have, since Trance(*e*) assigns probability 1 to all and only truths. If the only objection to adopting the trances as credence is that they are epistemically inaccessible (indeed, we are precisely as ignorant of the trances as of the future), we can accommodate ignorance of the trances in the same way we do with the chances by forming a mixture of trances obtained by conditionalizing on epistemically possible futures.

---

[38] See Ismael, "Raid! Dissolving the Big, Bad Bug."

[39] The arguments that $\mathrm{Pance}(e) = \mathrm{Pr}_G(e/\text{pre-2001 history of Australia})$ and $\mathrm{Fance}_t(e) = \mathrm{Pr}_G(e/\text{post-}t\text{ history})$ do not guide belief remain in place. Pance(*e*) would have us assigning crazy credences, and Fance(*e*) would have us assigning credences other than 1 (or 0) to events we know have occurred (or have failed to occur).

That is correct, so I need to say something a little more complicated than my earlier response. Notice that provided we use the same weights over epistemically possible futures in forming our best estimate of the trances and chances, our best estimate of future trances will be quantitatively indistinguishable from our best estimate of the chances. To say that our best estimate of the chances will be indistinguishable from our best estimate of the trances, however, is in its turn to say that in absence of information from the future, our best estimate of the chances is also our best estimate of the truth, since Trance(*e*) assigns probability 1 to all and only truths. So you can use the trances to guide credence if you want, but the calculation is a little more difficult. The chances and the trances come apart only in the presence of crystal balls.

This reinforces our explanation of the epistemic role of chance and tells us something interesting about the connection between chance and truth. Recall that PP said that provided that we have no magical information from the future, we should adopt the chances as our credences. We see now that we can derive a slightly more precise version of PP from the analytic and unqualified claim that truth should guide belief by adding a substantive, general claim about our epistemic position.

- If you know the truth, no matter what other information you have, adopt it as credence.
- The only specific information a situated human agent has about the world is information about past events.
- If you know the chances, provided you have no magical information from the future, adopt them as credence.

The lesson here is that chance is obtained from truth by forming a mixture over epistemically possible futures.

This also sheds some light on the question of what makes general probabilities conditional on world history more epistemically relevant than general probabilities conditional on background knowledge. Or perhaps better—why, if we have the latter, should we care about the former? The former are not more important than the latter, but they are prior in the sense that they provide materials used in calculations aimed at obtaining the latter. Many of the quantities we keep track of in physics have this character. They contain general, or generalizable, information that can be combined with locally available information to yield solutions to problems encountered in context. We can think of them as partially prepared solutions that get tailored for application by a relatively simple procedure. This strategy of partially preparing solutions is something we use every day to simplify frequently encountered problems. Look in your fridge for examples of partially assembled solutions to the problem of making dinner. Look at an electrician's toolbox for partially prepared solutions to the problems he encounters in his work. Or look in a mathematics textbook for useful theorems that play the same role for math students.

The second objection charges that we have reduced one form of probability to another but have made no progress toward explicit reduction of probabilistic facts to nonprobabilistic ones. To this, I reply as before that physics can, and in my view should, leave Pr_G unreduced, treating it on a par with other physical modalities. If one is after reduction, the prospects are much better following the Best-System Analysis in taking the whole theoretical package as the unit of reduction. But there are many (myself included) who deny that reduction is called for.

XII. CONCLUSION

Physics is and has always been awash with probabilities. Probabilities are not derivable from laws that constrain only possibility either in the deterministic or indeterministic case, and they are indispensible for purposes of prediction and confirmation in both. Physical laws that constrain only possibility do very little of the predictive work in practice. The same goes for confirmation. The existence of evidence that is logically incompatible with a theory is (almost) unheard of. All confirmation goes by way of probabilities; data disconfirms a hypothesis not by rendering it impossible but rendering it unlikely, and is only as objective as the probabilities invoked in comparisons of likelihood.[60] Something similar is true for explanation. There are always indefinitely many hypotheses that might explain a phenomenon. A *good* explanation is one that makes the phenomenon in question not physically necessary or inevitable, but *probable*. Once we recognize a form of general probability that is present in deterministic and indeterministic contexts alike, chance can be regarded as a form of single-case probability with a special epistemic role derived by conditionalizing on the past. This resolves the interpretive dilemma that treating chance as a fundamental form of probability generates, unifies classical and quantum probabilities, allows us to explain the role chance plays in guiding belief, and shows us how to see chance not as a new and fundamentally different form of probability but one rooted in a more familiar form of statistical probability.

---

[60] There are complex issues about how likelihood figures in confirmation, many of them unsettled, but there is no question that likelihood comparisons are indispensable in hypothesis testing.

There are some general lessons to be learned from this example. Chance, on my view, is an intermediate quantity, defined in terms of the fundamental ontology in a manner designed to play the role carved out by PP in guiding belief. It has an objective extension (if statistical probabilities are part of the mind-independent fabric of reality, then chances are too), but the special connection to credence is captured in PP. PP is not a binding principle of reason, something that holds for all rational creatures under all circumstances, but a principle of epistemic rationality that holds for creatures like *us*, for the *most* part, operating under quite specific constraints and facing quite specific tasks. We can imagine creatures faced with different tasks and operating under different constraints, or extraordinary situations into which we ourselves might be thrown, in which it would not make sense to adjust belief to chance. Chances would be equally present under those conditions, but they would not play the same role in guiding belief.

Chance is one example of a class of scientifically important structures that have this intermediate status. The interpretive problem for such quantities is to secure internal connections to perception and action while possessing objective truth conditions. Loewer, for example, characterizes the problem faced by interpretations of chance thus: "how could anything objective rationally constrain degrees of belief?" How could anything *other than* p *itself* that is part of the mind-independent fabric of reality rationally constrain degrees of belief in *p*? He argues that most influential accounts either fail to make chance objective or fail to secure the connection to belief.[61] Weslake and Price characterize the problem for accounts of the temporal arrow of causation and counterfactual dependence in similar terms. They say that such an account should not only identify the arrow with an objective feature of the world, but also explain its connection to action.[62] And, like Loewer, they argue that influential accounts fall on one or the other horn of this dilemma.

---

[61] "I survey the standard interpretations of probability...with an eye to evaluating whether any yields an a [*sic*] notion of chance capable of playing its role in physical theories and also as guiding belief via the PP My conclusion is that none of the usual suspects measure up. Those which interpret chance as an objective feature of fundamental events may account for its role in physics but make it utterly mysterious why chances so construed should guide our beliefs in conformity with the PP. Subjectivist accounts have no trouble with PP but are not credible as accounts of the chances that occur in physics." Loewer, "David Lewis's Humean Theory of Objective Chance," *Philosophy of Science*, LXXI, 5 (December 2004): 1115–25, at pp. 1117–18

[62] This is their Practical Relevance Constraint. In their words, any adequate account should "help us to make sense of a matter of great practical importance in our lives, the fact that we can act for future ends but not past ends (at least in normal circumstances)" Huw Price and Brad Weslake, "The Time-Asymmetry of Causation," in Helen Beebee, Christopher Hitchcock, and Peter Menzies, eds., *Oxford Handbook of Causation* (New York: Oxford, 2009), pp. 414–43, at p. 416.

Here, the interpretive strategy was to let the internal role of chance rigidly fix its extension, characterize that extension in objective terms, and then give a substantive account that adverts to contingencies of our epistemic and practical situation that explain why *those things* should play *that role*. I think this strategy can be employed usefully in other cases. I would argue that causal structure and other quantities with intrinsic temporal asymmetries, for example, can be understood in these terms.[38] The general viewpoint here is a pragmatic reorientation that recognizes that physical theories are practical and epistemic tools. They do not just serve as representations. They help us form beliefs about parts of space-time we do not have direct, independent sources of information about, and they help us intervene in nature effectively. They are not part of the fundamental ontology. They are best thought of as partially prepared solutions to frequently encountered problems. They are highlighted, brought into relief, and appear explicitly in formulations of our theories not because they have a distinguished ontological status, but because they play a special role in the practical and epistemic lives of creatures like us.

JENANN ISMAEL

University of Sydney and University of Arizona

---

[38] Huw Price has argued something similar; see his "Causal Perspectivalism," in Price and Richard Corry, eds., *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (New York: Oxford, 2007), pp. 250–92. It is his view, however, that these quantities are perspectival. That is not my view. Perspectival quantities are characterized by variation of truth conditions with contingent facts about their users or the epistemic or practical context in which they are employed. On my view, the truth conditions of the quantities in question are fixed and independent of facts about users or context of use. Discussion of these comes at the pragmatic level, in explanations of the role played by those quantities in our epistemic and practical lives

NEW BOOKS: ANTHOLOGIES

HELLER, MICHAEL and W. HUGH WOODIN, eds.: *Infinity: New Research Frontiers.* New York: Cambridge, 2011. xiii, 311 p. Cloth: $99.00.

HUTCHINGS, KIMBERLY and TUIJA PULKKINEN, eds.: *Hegel's Philosophy and Feminist Thought: Beyond Antigone?* New York: Palgrave MacMillian, 2010. xiii, 271 p. Cloth: $85.00.

HYDER, DAVID and HANS-JÖRG RHEINBERGER, eds.: *Science and the Life-World: Essays on Husserl's* Crisis of European Sciences. Stanford: University Press, 2010. xxxii, 247 p. Cloth: $60.00.

ILUNDÁIN-AGURRUZA, JESÚS and MICHAEL W. AUSTIN, eds.: *Cycling: A Philosophical Tour De Force.* Malden, MA: Wiley-Blackwell, 2010. xiii, 273 p. Paper: $19.95.

IRWIN, RUTH, ed.: *Climate Change and Philosophy: Transformational Possibilities.* New York: Continuum, 2010. xiv, 243 p. Cloth: $120.00.

JACQUETTE, DALE, ed.: *Cannibis: What Were We Just Talking About?* Malden, MA: Wiley-Blackwell, 2010. xix, 241 p. Paper: $19.95.

JAGGAR, ALISON M., ed.: *Thomas Pogge and His Critics* Malden, MA: Polity, 2010. viii, 272 p. Cloth: $69.95, Paper: $24.95.

JESHION, ROBIN, ed.: *New Essays on Singular Thought.* New York: Oxford, 2010. xii, 324 p. Cloth: $65.00.

JOLLEY, KELLY DEAN, ed.: *Wittgenstein: Key Concepts.* Durham, UK: Acumen, 2010. ix, 214 p. Cloth: $75.00, Paper: $24.95.

JOLLEY, MARC A. and EDMON L. ROWELL, JR., eds.: *Why Kierkegaard Matters: A Festschrift in Honor of Robert L. Perkins.* Macon, GA: Mercer UP, 2010. 265 p. Cloth: $45.00.

KAUFMAN, JAMES C. and ROBERT J. STERNBERG, eds.: *The Cambridge Handbook of Creativity.* New York: Cambridge, 2010. xv, 489 p. Cloth: $135.00, Paper: $65.00.

KELLY, MICHAEL R., ed.: *Bergson and Phenomenology.* New York: Palgrave MacMillian, 2010. xii, 277 p. Cloth: $85.00.

KOWALSKY, NATHAN, ed.: *Hunting: In Search of the Wild Life.* Malden, MA: Wiley-Blackwell, 2010. xvii, 258 p. Paper: $19.95.

KRAUSZ, MICHAEL, ed.: *Relativism: A Contemporary Anthology.* New York: Columbia, 2010. x, 573 p. Cloth: $99.50, Paper: $32.50.

KUL-WANT, CHRISTOPHER, ed.: *Philosophers on Art from Kant to the Postmodernists: A Critical Reader.* New York: Columbia, 2010. xi, 362 p. Cloth: $89.50, Paper: $29.50.

LEAR, JONATHAN and ALEX OLIVER.: *The Force of Argument: Essays in Honor of Timothy Smiley.* New York: Routledge, 2010. xii, 272 p. Cloth: $125.00.

LENNOX, JAMES G. and ROBERT BOLTON, eds.: *Being, Nature, and Life in Aristotle.* New York: Cambridge, 2010. xv, 289 p. Cloth: $99.00.

LOWE, SCOTT C., ed.: *Christmas: Better Than a Lump of Coal.* Malden, MA: Wiley-Blackwell, 2010. xii, 239 p. Paper: $19.95.

LYSAKER, JOHN T. and WILLIAM ROSSI, eds.: *Emerson and Thoreau: Figures of Friendship.* Bloomington: Indiana UP, 2010. xv, 203 p. Cloth: $60.00, Paper: $22.95.

MACDONALD, CYNTHIA and GRAHAM MACDONALD, eds.: *Emergence in Mind.* New York: Oxford, 2010. x, 286 p. Cloth: $74.00.

MACHAMER, PETER and GEREON WOLTERS, eds.: *Interpretation: Ways of Thinking about the Sciences and the Arts.* Pittsburgh: University Press, 2010. vii, 266 p. Cloth: $80.00.

MAGNUS, P. D. and JACOB BUSCH, eds.: *New Waves in Philosophy of Science.* New York: Palgrave MacMillan, 2010. xii, 274 p. Cloth: $100.00, Paper: $32.95.

MAKKREEL, RUDOLF A. and SEBASTIAN LUFT, eds.: *Neo-Kantianism in Contemporary Philosophy.* Bloomington: Indiana UP, 2010. vi, 331 p. Cloth: $70.00, Paper: $27.95.

MAYO, DEBORAH G. and ARIS SPANOS, eds.: *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science.* New York: Cambridge, 2010. xvii, 419 p. Cloth: $65.00, Paper: $36.99.

MCDONALD, CHRISTIE and STANLEY HOFFMAN, eds.: *Rousseau and Freedom.* New York: Cambridge, 2010. xiii, 312 p. Cloth: $85.00.

MCMAHON, JENNIFER L. and B. STEVE CSAKI, eds.: *The Philosophy of the Western.* Lexington: Kentucky UP, 2010. 368 p. Cloth: $35.00.

MCPHERRAN, MARK L., ed.: *Plato's* Republic*: A Critical Guide.* New York: Cambridge, 2011. xiii, 273 p. Cloth: $85.00.

MCQUILLAN, MARTIN and IKA WILLIS, eds.: *The Origins of Deconstruction.* New York: Palgrave Macmillan, 2010. x, 282 p. Cloth: $90.00.

MEISTER, CHAD, ed.: *The Oxford Handbook of Religious Diversity.* New York: Oxford, 2011. xii, 456 p. Cloth: $150.00.

MELAMED, YITZHAK Y. and MICHAEL A. ROSENTHAL, eds.: *Spinoza's* Theological-Political Treatise*: A Critical Guide.* New York: Cambridge, 2010. xiv, 294 p. Cloth: $85.00.

MENARY, RICHARD, ed.: *The Extended Mind.* Cambridge: MIT, 2010. 382 p. Cloth: $40.00.

MEYERS, CHRISTOPHER, ed.: *Journalism Ethics: A Philosophical Approach.* New York: Oxford, 2010. xxiv, 368 p. Cloth: $99.00, Paper: $24.95.

MONROE, DAVE, ed.: *Porn: How to Think with Kink.* Malden, MA: Wiley-Blackwell, 2010. xiv, 262 p. Paper: $19.95.

MORRIS, CHRISTOPHER W., ed.: *Amartya Sen.* New York: Cambridge, 2010. xvi, 224 p. Cloth: $92.00, Paper: $27.99.

MORRIS, KATHERINE J., ed.: *Sartre on the Body.* New York: Palgrave Macmillan, 2010. xii, 260 p. Cloth: $84.95.

MORRISON, DONALD R., ed.: *The Cambridge Companion to Socrates.* New York: Cambridge, 2010. xviii, 413 p. Cloth: $95.00, Paper: $29.99.

NANAY, BENCE, ed.: *Perceiving the World.* New York: Oxford, 2010. xii, 377 p. Cloth: $74.00.

NICOLÁS, JUAN ANTONIO, ed.: *Leibniz und die Entstehung der Modernität.* Stuttgart, Germany: Franz Steiner Verlag, 2010. 278 p. Paper: $59.00.

NIGHTINGALE, ANDREA and DAVID SEDLEY, eds.: *Ancient Models of Mind: Studies in Human and Divine Rationality.* New York: Cambridge, 2010. ix, 250 p. Cloth: $95.00.

NUDDS, MATTHEW and CASEY O'CALLAGHAN, eds.: *Sounds and Perception: New Philosophical Essays.* New York: Oxford, 2010. vii, 270 p. Cloth: $65.00.

CALL FOR PAPERS

THE TAYLOR CHARITABLE TRUST AND
THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL'S
PROGRAM IN THE HUMANITIES AND HUMAN VALUES
ARE PLEASED TO ANNOUNCE THE OPENING OF THE

First Annual

E. Maynard Adams Essay Contest

Prize Amount: $50,000.00

The prize will be awarded to the best review essay evaluating the merits of the ideas and arguments presented in E. Maynard Adams' *Ethical Naturalism and the Modern World-View* (1960). The late E. Maynard Adams (1919-2003) was Kenan Professor of Philosophy at the University of North Carolina at Chapel Hill.

Entries are sought primarily from tenure-track faculty in philosophy and religion, although other interested academic professionals are eligible.

Contest opens for registration Jan. 1, 2012. Participants must be registered by Mar. 31, 2012. Submissions must be uploaded to the contest website by Oct. 31, 2012.

For details and more information go to

www.emadams.unc.edu

(under revision, registration portal will open Dec 15, 2011)

ANNOUNCEMENT

Colin and Ailsa Turbayne International Berkeley Essay Prize Competition

The late Professor and Mrs. Colin Turbayne established an International Berkeley Essay Prize competition in cooperation with the Philosophy Department at the University of Rochester.

The next deadline for submitting papers is November 1, 2012. Submitted papers should address some aspect of Berkeley's philosophy. Essays should be new and unpublished and should be written in English and not exceed 5,000 words in length. All references to Berkeley should be to Luce/Jessop, and a MLA or similar standard for notes should be followed. Submissions are blind reviewed and will be judged by members of a review board selected by the Department of Philosophy at the University of Rochester. The winner will be announced March 1, 2013 and will receive a prize of $2,000. A copy of the winning essay will be sent to the George Berkeley Library Study Center located in Berkeley's home in Whitehall, Newport, RI.

Submissions can be sent electronically to: phladmin@philosophy.rochester.edu or by post mail to: Chair, Department of Philosophy, University of Rochester, P.O. Box 270078, Lattimore 532, Rochester, NY 14627-0078.

CAMBRIDGE

OUTSTANDING NEW TITLES *in* PHILOSOPHY

The Philosophy of Human Evolution

A Philosophical Introduction

MICHAEL RUSE

Cambridge Introductions to Philosophy and Biology

$85 00 Hb 978-0-521-11793-7
$27 99 Pb 978-0-521-13372-2
288 pp

Should Race Matter?

Unusual Answers to the Usual Questions

DAVID BOONIN

$99 00 Hb 978-0-521-76086-7
$34 99 Pb 978-0-521-14980-8
440 pp

Consciousness and the Self

New Essays

JEELOO LIU *and* JOHN PERRY

$90 00 Hb 978-1-107-00075-9
270 pp

Ethics and War

An Introduction

STEVEN P. LEE

Cambridge Applied Ethics

$95 00 Hb 978-0-521-89883-6
$29 99 Pb 978-0-521-72757-0
344 pp

Ethical Naturalism

Current Debates

SUSANA NUCCETELLI *and* GARY SEAY

$90 00 Hb 978-0-521-19242-2
270 pp

Globalization and Global Justice

Shrinking Distance, Expanding Obligations

NICOLE HASSOUN

$85 00 Hb 978-1-107-01030-7
220 pp

Kant: *Natural Science*

IMMANUEL KANT *and* ERIC WATKINS

The Cambridge Edition of the Works of Immanuel Kant

$130 00 Hb. 978-0-521-36394-5
1,110 pp.

Kant and Cosmopolitanism

The Philosophical Ideal of World Citizenship

PAULINE KLEINGELD

$90 00 Hb 978-0-521-76418-6
232 pp

Understanding Moral Obligation

Kant, Hegel, and Kierkegaard

ROBERT STERN

Modern European Philosophy

$90 00 Hb 978-1-107-01207-3
296 pp

Plato and the Divided Self

RACHEL BARNEY, TAD BRENNAN *and* CHARLES BRITTAIN

$99 00 Hb 978-0-521-89966-6
320 pp

Socrates' Daimonic Art

Love for Wisdom in Four Platonic Dialogues

ELIZABETH S. BELFIORE

$99 00 Hb 978-1-107-00758-1
350 pp

Descartes: Meditations on First Philosophy

A Latin-English Edition

JOHN COTTINGHAM

$90 00 Hb 978-0-521-19138-8
200 pp

The World-Time Parallel

Tense and Modality in Logic and Metaphysics

A A RINI *and* M. J. CRESSWELL

$95 00 Hb 978-1-107-01747-4
240 pp

Perception, Realism and the Problem of Reference

ATHANASIOS RAPTOPOULOS *and* PETER MACHAMER

$95 00 Hb 978-0-521-19877-6
300 pp

Third Edition!

Chinese Philosophy

HAIMING WEN

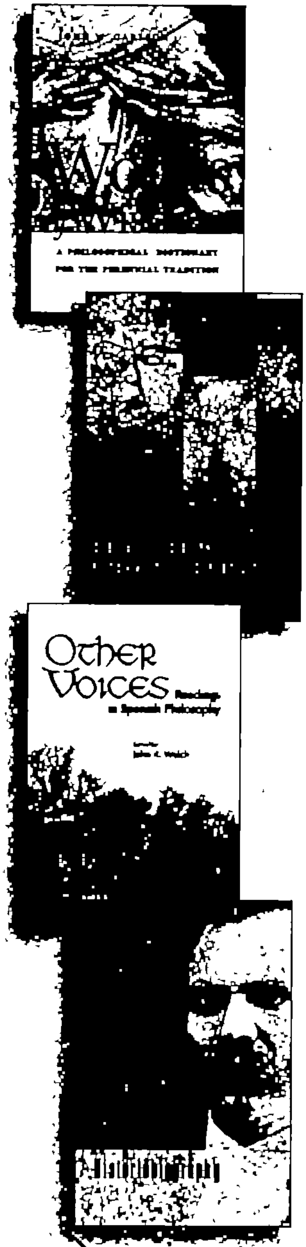
Introductions to Chinese Culture

$19 99 Pb 978-0-521-18676-6
152 pp

Prices subject to change

www.cambridge.org/us/philosophy

CAMBRIDGE
UNIVERSITY PRESS


Other Voices
Readings in Spanish Philosophy
A Philosophical Dictionary for the Perennial Tradition

Philosophy *from* Notre Dame

WORDS OF WISDOM
A Philosophical Dictionary
for the Perennial Tradition
John W. Carlson
978-0-268-02370-6 • $45.00 pa • 368 pp
E-book: 978-0-268-07693-1 • $29.00

THE VIEW FROM WITHIN
Normativity and the Limits of
Self-Criticism
Menachem Fisch and
Yitzhak Benbaji
978-0-268-02904-3 • $50.00 pa • 408 pp

NIETZSCHE AND THE
DRAMA OF
HISTORIOBIOGRAPHY
Roberto Alejandro
978-0-268-02037-8 • $40.00 pa • 392 pp

LEIBNIZ'S MILL
A Challenge to Materialism
Charles Landesman
978-0-268-03411-5 • $30.00 pa • 192 pp

VIRTUE AND POLITICS
Alasdair MacIntyre's Revolutionary
Aristotelianism
Edited by Paul Blackledge and
Kelvin Knight
978-0-268-02225-9 • $40.00 pa • 384 pp

OTHER VOICES
Readings in Spanish Philosophy
Edited by John R. Welch
978-0-268-04419-0 • $38.00 pa • 376 pp

ND UNIVERSITY OF NOTRE DAME PRESS
Web: undpress.nd.edu ◆ Tel: 800/621-27

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 9
SEPTEMBER 2011

page

447 *On Words* — John Hawthorne and Ernest Lepore

COMMENTS AND CRITICISM

486 *What Are Words? Comments on Kaplan (1990), on Hawthorne and Lepore, and on the Issue* — Sylvain Bromberger

504 *Words on Words* — David Kaplan

530 MEMO TO AUTHORS

531 NEW BOOKS: ANTHOLOGIES

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Teddy Seidenfeld, Gila Sher, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at www.journalofphilosophy.org

---

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U S copyright law All rights reserved. For more information about our copyright policies, please consult our January issue or contact us

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2011 by the Journal of Philosophy, Inc ISSN 0022-362X

UNIVERSITY
CALCUTTA

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 9, SEPTEMBER 2011

ON WORDS[*]

> Words strain,
> Crack and sometimes break, under the burden,
> Under the tension, slip, slide, perish,
> Decay with imprecision, will not stay in place,
> Will not stay still.
>
> —T. S. Eliot, *Quartet I, Burnt Norton*, 1935

In his seminal paper "Words,"[1] David Kaplan addresses a pair of questions that have been largely neglected by the philosophical community:

(i) Under what conditions are two utterances utterances of the same word?
(ii) What *are* words?

That these questions have not received much attention is rather surprising: after all, philosophers and linguists frequently appeal to considerations about word and sentence identity in connection with a variety of puzzles and problems that are foundational to the very

---

[*] Earlier drafts of this paper were presented at the Rutgers Semantics Workshop; the World Congress in Philosophy in Seoul, Korea, Saint Petersburg State University; University of California, Santa Cruz; University of Oslo; University of Paris; and the American Philosophical Association Pacific Meetings in April 2010. We would like to thank the audiences of each for their input, especially David Kaplan and Sylvain Bromberger. We also thank the anonymous referees of this JOURNAL for their helpful comments since this manuscript's original submission in June 2009.

We are most grateful to David Kaplan for his extensive, insightful, and witty commentary at the 2010 Pacific APA on sections I–III of our paper. We have made a few adjustments where we were convinced that we had not paid sufficient attention to his original text. We have flagged a few other places where his public remarks gave us additional evidence concerning his original intentions. We have tried for the most part to make the adjustments in footnotes, in particular, we added or altered footnotes 12, 20, 22, 26, 29, and 37.

[1] David Kaplan, "Words," *Aristotelian Society Supplementary Volume*, LXIV (1990): 93–119. Subsequent references to this work appear in the text.

0022-362X/11/0809/447–485

subject matter of philosophy of language and linguistics.[2] Kaplan's attention to words is thus to be applauded. And there is no doubt that his discussion contains many useful insights. Nevertheless, we find his picture deeply flawed for a variety of cross-cutting reasons. Our aim in this paper is to further advance an understanding of the nature of words, both by remedying the problems with Kaplan's account, and also by achieving a suitable perspective on what the metaphysical investigation of word identity can hope to achieve.

Our discussion divides into four sections. In section I, we evaluate Kaplan's discussion of a contrast integral to his own account: that between the type-token and the stage-continuant conceptions of words. In section II, we present three constraints on any account of words and two further themes in Kaplan's discussion central to his conception of words—the role of repetition and the constitutive authority of intentions. While these ideas have laudable motivations, we argue they are far from the best way of making good on the insights that drive them. The final two sections take a skeptical turn. In section III, we express doubt about Kaplan's presumption of the importance of what he calls 'common currency names', thus raising a suspicion that he may be in pursuit of a chimera. Finally, in section IV, we express pessimism about whether interesting answers to question (i) above will be forthcoming. Does the legitimacy of our word ontology need or depend on the availability of such answers? We explore these issues in some detail, mindful of the fact that the case of words is in many ways representative of many other ontological categories from the manifest image. Along the way, we tease apart a number of metaphysical questions in the vicinity of the topic of word individuation—questions rarely disentangled—and consider how the discussion of the previous parts bears on them.

I. TYPE-TOKEN VERSUS STAGE-CONTINUANT

There are physical events in spoken language—utterances. There are physical products of written language—inscriptions. There are physical events in sign language—signings. And there are words that certain utterances, inscriptions, and signings are utterances, inscriptions, and signings of. What exactly is the relation between the former class of events and objects and the words themselves? Kaplan presents us with two candidate models.

---

[2] See, for example, the discussions of Motivations Three through Five in section III below. The importance of word individuation to mainstream philosophy is not new; notice, for example, its relevance to those varieties of nominalism that analyze the phenomenon of having something in common in terms of *the same word* being truly predicable of many objects.

On the first model of words, the *type-token model*, a word is an abstract object instantiated by various physical objects and events—*tokens*. Further, as Kaplan conceives it, the type-token model treats the abstract object as encoding a certain form that is common to the tokens. Tokens thus get to be tokens of a word by having a physical embodiment with the right intrinsic profile. On the simplest and most common view, word identity begins with orthographic and/or phonetic shape,[3] that is, with spelling and pronunciation. Two words are identical just in case they share the same relevant shape.[4]

On the orthographic side, Wetzel tells us that "'dog' is a different expression from 'god'. Not only the letters, but the *order* of the letters occurring in an expression determines the expression's identity."[5] Orthographically, then, a word is identified with a sequence of letters with a space at each end and no space in its middle.[6]

Linguists, unlike philosophers and lexicographers, ignore orthography altogether and concentrate on sound, in particular, on sequences of phonemes. When Fiengo and May say, "...phonological distinctness, when it is believed to fall outside the range of dialectal variation, may be individuative of words,"[7] they intimate a partial discrimination of words through sound, as does Bromberger when he writes of the *word* 'dog' that it "is monosyllabic."[8] Segal and Speas reject Davidson's claim that the 'that' of indirect quotation is the same word as the orthographically indistinguishable demonstrative word 'that' by noting that their phonologies are distinct.[9]

---

[3] Donald Davidson, "Quotation," in *Inquiries into Truth and Interpretation* (New York. Oxford, 1979), pp 79–92, at p. 90 Compare with Susan Haack, *Philosophy of Logic* (New York: Cambridge, 1978), p. 75: "[Expressions are] either a pattern which similar tokens exemplify, or a class of similar tokens"; and Philip Hugly and Charles Sayward, "Expressions and Tokens," *Analysis*, XLI, 4 (October 1981): 181–87, at p 184: "expressions...can be construed as classes of perceptible particulars similar in some physical respect to given perceptible particulars "

[4] "An expression is a phonic or graphic form· the former has acoustic properties and the later spatial ones." See Brendan S Gillon, "Ambiguity, Indeterminacy, Deixis, and Vagueness: Evidence and Theory," in Steven Davis and Gillon, eds., *Semantics: A Reader* (New York Oxford, 2004), pp. 157–87, at p. 162. Compare with John Lyons, *Semantics: Volume 2* (New York: Cambridge, 1977), p. 18.

[5] Linda Wetzel, "What Are Occurrences of Expressions?" *Journal of Philosophical Logic*, XXII, 2 (April 1993): 215–19, at p. 218.

[6] W. V O. Quine, *Word and Object* (Cambridge: MIT, 1960) recognized that "the principle behind the printer's use of spaces is dim," p. 13. What do we make of expressions like 'ice cream' and 'ice-cream' or, for that matter, languages with writing systems without spaces—as with most ancient languages, including Phoenician, Greek, and Latin, and Medieval Latin? Yet—as do most philosophers—Quine acquiesces to the orthographic account (*ibid.*, p 14).

[7] Robert Fiengo and Robert May, *De Lingua Belief* (Cambridge: MIT, 2006), p. 54.

[8] Sylvain Bromberger, "Types and Tokens in Linguistics," in Alexander George, ed., *Reflections on Chomsky* (Cambridge: Blackwell, 1989), p 85.

Concentration on pronunciation as the identifying feature of words extends back to Aristotle, who treated written words as mere 'stand-ins' (or 'subordinates' (Ockham)) for spoken ones. However, some have noted (Pierre d'Ally) that a written message can be understood directly, without recourse to sound; there are so many words we know how to spell but not pronounce.

According to either orthodoxy, then, nothing in a word weds it to any particular grammatical category, meaning, or even a specific language. The same word (that is, sounds or shape) might be both a noun and a verb, carry several or even no meanings, co-exist in different languages, or change any of these features across time and location.[10]

On the second model of words, the *stage-continuant model*, which Kaplan prefers, "utterances and inscriptions are *stages* of words, which are the *continuants* made up of these interpersonal stages" (98). The phrase 'made up' implies a compositional relation. The picture conveyed is that words are long-lived objects composed out of short-lived objects—utterances and inscriptions. Given these remarks, it is natural to elucidate the stage-continuant model using a framework that is standard to four-dimensionalist metaphysics: a word is a four-dimensional continuant with various utterances or inscriptions as its short-lived stages.[11] A particular utterance or inscription of a word *w* thus turns out to be a spatiotemporal part of *w*: *w* is a fusion of the archipelago of its spatiotemporal parts. Of course, as Kaplan realizes, this still leaves many questions unanswered.[12]

The stage-continuant conception does not, in and of itself, tell us when a fusion of utterances or inscriptions counts as a word, and it does not tell us when two utterances or inscriptions are parts of the same word continuant. But this is as it should be. By analogy, the stage-continuant model of personal identity, according to which people are fusions of person stages, does not tell us under which conditions two stages belong to a single person continuant and under which conditions a fusion of person stages counts as a single person. But both at least purport to provide a framework within which the metaphysics of persons and words can be suitably pursued.

---

[9] Gabriel Segal and Margaret Speas, "On Saying That," *Mind and Language*, I, 2 (June 1986): 124–32.

[10] Compare with Simon Blackburn, *Spreading the Word* (New York: Oxford, 1984), p. 20.

[11] Truths to the effect that a certain word undergoes change will be accommodated by such a conception by standard four-dimensionalist accounts of what change over time comes to.

[12] Kaplan's APA presentation indicated that while, to our mind, his written exposition suggests the four-dimensional conception, he did not really intend to be endorsing a metaphysical picture of that sort. His presentation also made clear that he intended the stages of words to include the internal token representations by means of which a word is stored (even though his actual gloss on the nature of word stages does not make this intention manifest).

First, we wish to emphasize that we quite agree with Kaplan that a philosophically satisfying theory of words cannot proceed entirely within a shape- or form-theoretic framework. Simple reflection on the fact that the same word can be written, uttered, signed,[13] Brailled, or semaphored already renders such a conception dubious: after all, there is hardly anything shape- or form-like in common among an utterance, an inscription, a hand gesture, and a bump on a panel. In sum, the form-theoretic conception tends to ignore the fact that a variety of systems can articulate a single word. As Kaplan emphasizes (see, in particular, 102–06), even if we restrict attention to a particular system of articulation, the form-driven conception does not square with the fact that the way a word is articulated within that system can change *its* shape or form significantly over time and location. The word 'color' has two "modern current or most usual spellings" ('colour', 'color') and eighteen historical ones.[14] Accordingly, 'color' and 'colour' are two spellings of the same word.

We do not deny that there may be *certain* contexts in which we use 'word' to speak about entities that are individuated by shape or form. After all, there are contexts in which we can communicate truths with such things as "The word 'color' has exactly five letters," "The words in that book are a funny shade of gray," and so on. This kind of flexibility is not particularly surprising. Consider the flexibility of our uses of 'that book': sometimes we use it to indicate a particular token volume; sometimes a particular work, say, Dante's *Inferno*; sometimes a particular edition ("I wish I owned Aldus's edition of Dante. I have two friends that own that book.").

The form- or shape-theoretic conception *may* capture some of our uses of 'word' (we shall return to this topic below). But there *are* clearly uses for which nothing like that conception is adequate. Like Kaplan, we are interested in these latter uses.

(The form-theoretic conception is no doubt driven by the insight that a word's meaning is inessential to it. Crudely put, the key idea is that if the meaning is inessential, then what is left except form? As Kaplan's own discussion renders clear, though, this thought turns on a failure to realize that relational factors may be crucial as to whether an utterance or inscription is an utterance or inscription of a word. We return to this issue in due course.[15])

---

[13] We have in mind Signed English, a system of manual communication that strives to be an articulation of English vocabulary and grammar, in contrast to American Sign Language, a separate language with its own vocabulary and grammar.

[14] *Oxford English Dictionary*, 1971 ed., vol. 2, s. v. "color"

The second—more critical—point we wish to register is that Kaplan presents us with a false dichotomy. (We concur with McCollough, though we disagree with him that all tokens of a word must be "similar or resemble" each other in an interesting intrinsic respect.[16]) One can favor the type-token conception in thinking of a word as an abstract object without taking on board any commitment to a shape- or form-theoretic conception of words. Consider, by analogy, species: in repudiating a view of species as four-dimensional objects, one need hardly be committed to a shape- or form-theoretic view of what species are. The bare-bones, abstract-object view of species merely tells us that a species is a nonconcrete object and that there is a relation which holds between a particular creature and its species—the *of*-relation. Nothing requires that the conditions for the of-relation must hold in virtue of the intrinsic features of the creatures. In particular, it is quite compatible with the bare-bones view that historical-relational features are crucial to whether an individual belongs to a particular species.[17]

---

[15] In a classic study, Peter Ladefoged and D. E. Broadbent manipulated the formant frequencies of material preceding a target word that could be perceived as 'bit', 'bet', 'bat', or 'but'. These manipulations to the material affected the way listeners categorized the final target word, producing context effects such that the target word was identified differently as the formant frequencies of the preceding material changed. The spectral manipulations to formant frequencies of the precursor material created a perceived change in voice characteristics, making it sound like different speakers uttered the target. See Ladefoged and Broadbent, "Information Conveyed by Vowels," *Journal of the Acoustical Society of America*, XXIX, 1 (January 1957): 98–104.

We interpret this experiment as evidence against the view that the intrinsic character of an utterance determines whether it is an utterance of a certain word. As we mention in the text, there may be uses of 'word' which are attuned to form. These reflections show that if such conceptions are not bankrupt, founded on illusions of commonalities that are not present, then even they cannot be explicated by appeal to intrinsic character alone. It is also not hard to make trouble for an orthographic conception that presumes that, in the relevant sense, orthographic form is intrinsically determined. Note, for example, that a physical space appears between the articulated letters of a single word even though these do not realize a space in the orthographic sense. Our judgment as to whether a physical space constitutes a space in the orthographic sense depends upon the relative size of that space to other spaces. If the gap between the letters 't' and 'o' is half an inch but flanked by spaces of two inches, we are not deterred from treating the sequence 't' and 'o' as expressing a single word. This all makes further trouble for those who think that the intrinsic character of an expression determines whether or not it articulates a word, since the relative-size facts just adverted to are not intrinsic to the expression itself.

[16] Gregory McCulloch, "Making Sense of Words," *Analysis*, LI, 2 (March 1991): 73–79, at p. 74.

[17] Kaplan, *op. cit.*, p. 98 seems to think the stage-continuant model is required in order to accommodate historical-relational insights. But he gives no justification for this contention

We conclude that even when the words-as-forms view is repudiated, a real choice of metaphysical framework remains: we might, in line with the picture suggested by Kaplan's stage-continuant gloss, think of words as four-dimensional objects which have inscriptions or utterances as temporal parts.[18] Or, we might just as well treat a word as a nonconcrete object which is articulated by various concrete events or objects. We shall call this the *abstracta-articulations model*.[19,20]

We have two general reasons for preferring the abstracta-articulations model over the stage-continuant model. Our first concern with the stage-continuant model turns on a more general choice point of modal metaphysics. It seems clear enough that various facts about the performance profile of a word are *inessential* to the word. Uses of a given word could have become obsolete. It once might have been articulated more loudly; and so on. Supposing a word to be an archipelago of utterances and inscriptions, how are these sundry facts to be accommodated? One particular source of disquiet is that insofar as we think of the continuant as an archipelago of utterances, our modal intuitions may not at all match the modal profile with which the word was originally associated. This concern is familiar from discussions of four-dimensional objects in general: supposing that a person is a fusion of person stages, and given that there are ways of thinking about aggregates of stages which yield very different modal intuitions, how do we accommodate such facts as that a person could have lived for a much shorter or longer time?[21] The standard mode of accommodation in person-as-continuant metaphysics is via counterpart theory: various possible fusions stand in the person-counterpart relation to the actual continuant, and various ascriptions of essential and accidental properties to the continuant are rendered true or false by the profiles of the possible fusions that are person counterparts of that continuant.[22]

---

[18] We note that one might wish to allow that one temporal part of an inscription is a temporal part of one word and a later temporal part is not part of that word—just as an earlier but not later temporal part of an atom might be part of a person.

[19] We ourselves would have no deep objection to using 'type-token' language for this model, but since Kaplan insists on a shape-theoretic construal of the language of types and tokens we shall stay clear of it.

[20] There are yet other models that might be considered. As Mark Johnston emphasized to us (personal communication), one might opt for the view that words (and species) are higher-order persisting objects that are concrete; that are to be sharply distinguished from their realizations in utterances, inscriptions, and so on; and that—*contra* the stage-continuant model—are best thought of within the framework of three-dimensionalism. On this view, as against the abstracta-articulations model, words only exist at a world insofar as some first-order realization of them does. At no time can a word exist without some physical object constituting it. Meanwhile, as against the stage-continuant model, words should not, on this view, be thought of as composed of the multitude of their tokens. Rather, at any particular time there will be certain tokens that constitute it. As will be clear from our discussion, our own current preference is for the abstracta-articulations model. But we recognize that the space of positions ought not to be limited to those described in the main body of the text. In general, those who wish to think of words as concrete will have more intellectual wiggle room insofar as they allow that two concrete objects can occupy the same four-dimensional region of space-time. We lack the space to pursue these issues further here. (We note that in discussion Kaplan expressed a preference for something like Mark Johnston's view, as opposed to the stage-continuant view on its four-dimensionalist construal. We also acknowledge the need for further clarification of what the purported three-dimensionalism/four-dimensionalism contrast comes to.) For more, see John Hawthorne, "Three-Dimensionalism," in *Metaphysical Essays* (New York: Oxford, 2006), pp. 85–110.

We anticipate that insofar as one pursues a continuant model of words, one will likely be drawn into an analogous account of attributing modal ascriptions to words. Crucially, however, the viability of the approach will turn on the viability of the counterpart-theoretic approach to *de re* modal judgments. While this is not the place to pursue these more general issues, our inclination is to be wary of counterpart theory and, for that reason, to be wary of the stage-continuant model. Insofar as a reader shares our suspicion of counterpart theory, he should at least be sensitive to issues as to how the stage-continuant model is supposed to accommodate modal platitudes without immersing itself in an objectionable modal meta-semantics.[23]

Our second concern with the stage-continuant model of words is more straightforward and ought to have purchase whether or not one is a fan of counterpart theory. It is clear enough that Kaplan is committed to the view that all words are articulated in one way or another. ("They live in the world, not in Plato's heaven" (111).) If a word is an archipelago of articulations, then absent a single articulation there is *no* word. But this is an untoward result. Consider, for example, the fact that the prefix 'un-' ('im-', 'in-', 'il-') can combine with ever so many adjectives to produce a word.[24] It seems evident that a particular combination of that prefix and an adjective—as in 'unhappy', for example—might never have been articulated.[25] But that would hardly preclude this combination from being a word of the language. It may even be that in this scenario many people thought about using the word 'unhappy' but decided against it, or intended to use it but never got around to it.[26] Given this presumed profile of thoughts and intentions, it would be very awkward to deny that the word exists. And yet it is very hard to see how it could exist once a Kaplanian framework is adopted. Indeed, the case could be made stronger; once one realizes that some morphology is productive, as in 'anti-missile', 'anti-anti-missile', and so on, it becomes clear that there are more unspoken words than spoken ones (*contra* Kaplan, who insists, "The world is not brimming with unspoken words" (117)).

---

[21] A standard puzzle in the vicinity is the following Take the worm that is a certain person x and the worm that is the temporal part p of x that spans the first thirty years of x's life. Suppose x lives sixty years It is possible that x could have lived only thirty years and had a life that matched x's actual life. It is natural to think that at such possible worlds p exists. By hypothesis, at such worlds x exists. But assuming the classical extensional mereology standard among four-dimensionalists, we are then under pressure to say that x and p are identical at those worlds, violating the necessity of distinctness. For relevant discussion, see Peter Van Inwagen, "Four-Dimensional Objects," *Noûs*, XXIV, 2 (April 1990): 245–55. A second puzzle is this: when we think of a person as a concertina of stages, it is *not* natural to think that it could have been very short in temporal duration, while when we think of the person as a person, it is natural Which perspective is to be privileged? For relevant discussion, see David Lewis, "Counterparts of Persons and Their Bodies," this JOURNAL, XLVIII, 7 (Apr. 8, 1971): 203–11. The same puzzles arise in the current setting, *mutatis mutandis*. Counterpart theory offers perhaps the most promising way of navigating such puzzles to advocates of temporal parts who are wedded to standard mereology.

[22] Obviously, counterpart theory is not Kaplan's preferred way of thinking about modality. The point is that the picture suggested by the thesis that words are 'made up' out of their stages sits best with counterpart theory.

[23] See Michael Fara and Timothy Williamson, "Counterparts and Actuality," *Mind*, CXIV, 453 (January 2005): 1–30.

The Kaplanian might try to retreat by claiming that the stage-continuant model is only appropriate for words that are not built up out of other words, and propose a different account of the remainder.[27] But this is inadvisable. For one thing, we surely would prefer a more unified account if we could get it. Second, such a retreat fails to appreciate the import of the word-morpheme distinction. 'Un-' is not a word; it is a morpheme.[28] In the imagined scenario, the unarticulated existence of 'unhappy' as a word cannot be explained via the existence of two simple words, 'un-' and 'happy', since this would misclassify 'un-'. (The same can be said for the morpheme 'anti-' in the word 'anti-missile'.)

---

[24] Gillon, *op. cit.*, p. 164.

[25] One might object, as does Jeff King (personal communication), that since morphology is not fully systematic—it is not in general true that any combination of 'un-' with an adjective yields a word—it makes no sense to suppose that 'unhappy' is a word and yet lacks concrete realization. Note that there are languages where morphology is much more systematic (Mohawk is one; thanks to Mark Baker), and hence, where this objection would not have purchase. This worry has less force against certain other examples from English—see our 'anti-anti-missile' example below. In that case, the application of a rule for 'anti-' is clear even in the absence of the relevant concrete instance. Note also—deploying a theme from the main text—that one could stipulate that 'un-' does combine with 'happy' without actually concretely realizing 'unhappy'. In that case, the anticipated complaint has much less force.

[26] Of course, if internal storage counts as articulation (cf. note 12), then a time when people thought about using 'unhappy' may, on a natural construal, count as a time when it was articulated, albeit internally. While this appeal may help the stage-continuant view in some cases, it will not provide the resources to block the unwelcome conclusion that, at times when the particular combination of morphemes is not even internally articulated, the relevant word passes out of existence. Moreover, if one thinks that when a word stops being articulated it stops existing and *is incapable of coming back into existence*—in conversation, it became clear that this is an important part of Kaplan's anti-Platonism—then one risks having to say that a new word 'unhappy' gets considered whenever there is a lapse during which the relevant combination of morphemes never gets articulated.

[27] As far as we can see, there is no anticipation of such a retreat in the original paper.

It is *also* perfectly possible that something *un*articulated counts as a word because it is an acceptable combination of morphemes, none of which is itself a word. But the anticipated retreat lacks the resources to accommodate this scenario. To reinforce this point, one need only consider possible facts about word invention. Suppose the impositor proclaims, "The sequence of morphemes 'un-' followed by 'voke' is to mean such and such." It is intuitively clear that his speech serves to *introduce* the word 'unvoke' into the language. Suppose soon after his decision, the world blows up. The word 'unvoke' has been introduced by a description and so has become a constituent of the language, but it has never been *used* in discourse. Notice, in particular, it did *not* occur in the speech that served to introduce it.[29]

---

[28] Words are distinct from morphemes. They are the smallest units of a sentence with positional mobility, meaning roughly, they are the smallest units of a sentence redistributable *salva congruitate*. The words 'men' and 'women', for example, display their distributional freedom in (a)–(b):

(a) Do men like women?
(b) women men like

Morphemes, on the other hand, have a rather fixed sequential order, as evidenced in (c)–(d).

(c) unhappy
(d) *happyun

Syntax can manipulate words but not smaller units. As a corollary, it is true that there are 'free word-order' languages (to varying degrees, including English), but no 'free morpheme-order' languages.

Morphemes, unlike words (and complex expressions), resist linguistic material insertion. (a) permits the insertion of an indefinite number of new items, as in (e)–(f):

(e) Do men who are single like women who are single?
(f) Do men often like women?

More generally, words can be inserted between words, and morphemes can be inserted between morphemes (when conditions are right), but words cannot be inserted between morphemes. For example, we say of (g) that 'the' is a word and '-s' is a morpheme partly because we can insert words between 'the' and 'boy' but not between 'boy' and '-s', as in (h) and (i):

(g) the boys
(h) the tall boys
(i) *the boy responsible s (though 'the boys responsible' is fine)

Another large issue here is phonological: words are self-contained phonological units; morphemes are not. Words have at least one syllable, their own stress, and so on. Morphemes need not (see again plural '-s'). (Thanks to Mark Baker.)

The abstracta-articulations model faces no problems in accommodating word-introduction scenarios of this sort. But they embarrass the stage-continuant model, since there are no stages upon which to ground the existence of the word. (We similarly can introduce a written word descriptively by saying which sequence of letters comprises it. If we say, "Let the sequence of letters 'L' followed by 'U' followed by 'V' mean *love*," though we have introduced a word, we have not yet tokened it.)

Of course, the abstracta-articulations model for words is logically compatible with a stage-continuant model for morphemes. Might the Kaplanian theory be salvaged by reconfiguring it as a view about morphemes? Such a view handles some of the problems registered in the text, but not the last one raised in the above paragraph.[30]

(Consideration of complex words also suggests the following, admittedly more recherché, worry for the stage-continuant model. There could be a language where one constructs a word w_2 by placing a word w_1 after itself, and where the only uses of w_1 are as parts of an articulation of w_2. For example, it might have been that

---

[29] Again, matters become more delicate if we allow articulations to include internal articulations. For in the situation described, it is natural to think that the word 'unvoke' is internally represented. In his APA presentation, Kaplan wished to (a) require that there be such an internal articulation in order for the word to come into existence and (b) insist that once all articulation lapses then the word permanently goes out of existence. We think that there is a strong case against (b). Suppose we introduce 'unvoke', use it for a while, and then forget about it. No articulation remains. A few years later we come across a video that depicts the original introduction speech. It captures the moment where the speech "Let the sequence 'un' followed by 'voke' be a word that means " is made. Having witnessed the video we pick up using the word again. (b) will reckon such a description of the case incoherent: one will be forced to think of a new word coming into being when the video is witnessed. Notice it is no defense to say that since the video was in existence all along there was an articulation of the word all along. This is to strain the notion of articulation beyond any useful sense. Given that the original overt description did not articulate the word, it is clear that in no reasonable sense does the video of that speech do so. (We note in passing that it is far from clear in any case that the picture that words are concrete changeable objects ought itself to deliver (b). One might think that the very same boat or watch may exist at two times in between which there is a time when it does not concretely exist owing to its being disassembled for winter storage or for repair by the watchmaker. And surely no one needs words to be more concrete than boats and watches.) One test case for (a) is where the structure of internal articulation matches that of the overt speech: there is an internal description but in no reasonable sense an internal realization of the word. We shall not explore this or related test cases further here.

[30] Note that if Kaplan's view were reconfigured as a view about morphemes, it might put pressure on what he says about the constitutive role of intentions, since ordinary intentions to repeat do not seem to have morphemes as their objects.

'people' was only ever used in constructions like "They are people people." Perhaps all sorts of other uses are allowed, but as a brute contingent fact that potentiality is never actualized. In this scenario, a fusion of articulations of w_2 would appear to have exactly the same four-dimensional profile as a fusion of articulations of w_1. If one embraces a unique-fusion principle for continuant mereology, one will be forced to admit the untoward result that w_1 and w_2 are the same word!)[51,52]

---

[51] The unique-fusion principle says that at most one thing fuses a given class of objects. The intuitive idea is that one disallows two objects made of the same bits.

[52] Given that morphemes are abstract objects and that the two putative occurrences seem to involve a numerically identical morpheme, what does it mean, exactly, to say one morpheme occurs *twice* in a word ('anti-anti-pacifist', for example)? We will make a few observations in connection with this constituency question:

(i) Do *not* think of words as built up out of letters. That conception of words privileges written articulations of words in an unjustified way, since there is no good reason to think of spoken words as built up out of letters.

(ii) The problem should be treated on all fours with a variety of intimately related problems: what does it mean to say that the proposition *that two is even* 'occurs twice' in the disjunctive proposition *that two is even or two is even*? What does it mean to say that there are two hydrogen molecules in H_2O? In each case it is not clear how to resolve the issue with a mundane appeal to type-token language, since the relevant constituents seem to be abstract objects rather than 'tokens' of them. Issues of this sort have frequently been discussed by metaphysicians under the heading 'structural universals', and we cannot do full justice to the range of issues involved here. But we can gesture at our own preferred solution.

Often, what postures as counting of a certain sort of object is, at bottom, a mode of counting a number of states and events that involve that object. Thus, recalling a standard example, we may say that twelve barges came through a lock on a particular canal on a particular day, even where this involves three physical barges passing through four times each. What 'twelve' is a count of is not some physical barges, but instead, plausibly, of twelve barge-involving events (in particular, events of a barge passing through that lock). Consider now the singular proposition *that John loves John*. The claim that John occurs twice in the proposition *that John loves John* is even less problematic. Even on its face it parades as a count of events. With a suitably broad conception of events we can identify two such events: the event of John, the individual, being the agent of the proposition, and the event of that very same individual being the patient of the proposition. Same individual, different events. There seems to be no deep obstacle to extending this model to the case at hand: the two occurrences of the morpheme 'anti-' are two events involving one and the same morpheme.

(iii) Even supposing one can make good on these ideas, a number of residual mereological queries remain. Is there anything that overlaps the word 'anti-anti-missile' (where x overlaps y iff x and y share a part) but which does not overlap 'anti-missile'? If you answer 'no', then you have to depart from standard mereology, which tells us that if x is not y, then there is something that overlaps one but not the other. If you answer 'yes', then you have to find exotic objects that are suitable mereological difference makers. These decisions—intimately related to David Lewis's suspicions about 'structural universals'—are difficult, but they are not special to the mereology of words. They apply with equal force to other examples involving pairs of objects which are, roughly speaking, alike in matter but not form. Resolution of these questions is obviously beyond the scope of this paper.

II. THREE CONSTRAINTS AND TWO PROBLEMATIC IDEAS

Much of what Kaplan says about the stage-continuant model carries over to the abstracta-articulations model. Hence, the discussion in section I leaves many evaluative questions about Kaplan's discussion unresolved. In this section, we first identify a number of themes that shape his discussion and that we shall also adopt as framing principles. Second, we identify two of his central ideas that, even assuming those framing principles, strike us as seriously problematic, and we suggest how they can be improved upon.

One of Kaplan's guiding insights concerns the evolution of words: over time there may be considerable variation in how a word is written or pronounced. As it is transmitted from one epoch to another, it can change not only its phonological and orthographic contours, but its semantic and grammatical ones as well. Like Kaplan, we are after a metaphysics of words that satisfies this *Evolutionary Constraint*.

A second idea that is certainly implicit in Kaplan's account is that a word can be articulated in untold different systems. One and the same word can be written on a pad with a pen, typed on a sheet of paper, projected on a screen, spoken out loud, signed with a gesture, and Brailled on a plaque. Different media can be used to express the same message. Only our imaginations limit how we go about articulating words. Call this the *Multiplicity Constraint*.[33]

A third idea we shall try to respect is that two utterances or inscriptions, even within a particular community, may in an important sense of the word 'word' count as utterances or inscriptions of different words even though they are exactly alike in how they sound (in the case of two utterances) or in how they are written (in the case of inscriptions). Call this the *Coincidence Constraint*.

In addition to these three constraints, two further themes are central to Kaplan's conception of words and word performances. The first is his commitment to the importance of the concept of *repetition*. ("This notion of repetition is central to my conception" (103).) Suppose someone articulates a word in a performance. What makes a subsequent performance a performance of that word? (The question can be posed—though need not be—within the stage-continuant model: suppose someone produces a stage of a word continuant. What makes a subsequent performance a stage of the same continuant?[34])

---

[33] There may be (contingent) priority relations between certain pairs of articulation systems. Thus one might argue that the alphabet works by representing phonology, rendering the former parasitic on the latter. We shall not enter into these priority disputes here.

[34] "What is it that makes a particular output, the transmission of the same word as that carried by a particular earlier input?" Kaplan, *op. cit.*, p. 102.

In answering this question, Kaplan is responsive to the fact that there need not be an intrinsic match between the performance of a speaker and a prior performance he has encountered in order for both to be performances of the same word: there may be significant differences in shape and sound. (To reinforce this point, notice that a speaker may encounter a spoken performance and then go on to produce a written one.) As a remedy to any presumed significance of an intrinsic match, Kaplan stresses in its place the importance of *intentional repetition*: what is important is that a speaker intends to *repeat* a word. This suggests the following sufficient condition for two performances to be of the same common currency name, which we shall call *Repetition*:

> *Repetition*: If p_1 is a previously encountered performance of a word w and p_2 is performed in such a way that its utterer intends p_2 to be of the same word as p_1, that suffices to establish that p_2 is a performance of w.

This sufficient condition hardly vindicates the importance of repetition, since it is compatible with Repetition that word matching is almost always secured by a different mechanism.[35] Kaplan is evidently committed to the pervasiveness of the repetition mechanism. What is the cash value of the requirement that repetition be pervasive? Given that, to our ear, it is not natural to speak of repeating except where some performance is being repeated, an instance of the pervasiveness of repetitive intentions suggests the following thesis:

> *Pervasiveness*: For any noncreative performance p_1 of a word w, there has to be a previously encountered performance p_2 of w such that p_1 is produced with the intention of repeating p_2.

The further central theme to which Kaplan is apparently committed is implicit in Repetition. Roughly speaking, the theme is *intending it to be so makes it so*, which we shall call the *Constitutive Role of Intention*:[36]

> *Constitutive Role of Intention*: If someone intends to produce the same word w as that used in a particular performance, then whatever comes out of his mouth (or from his pen) is a performance of w.

---

[35] Ruth Millikan, *Language, Thought, and Other Biological Categories* (Cambridge. MIT, 1984), p. 72 also ties word identity to a history of uses when she says, "the Martian who due to historical accident utters what sounds just like the French 'Il pleut,' even though he may happen to mean just what the Frenchman typically means when he uses this sound, does not utter the French but rather the Martian word 'Il'." But she minimizes the role of intention, "the parrot that mimics tokens of the English word 'hello' tokens *the English word 'hello'* and the child who uncomprehendingly copies out 'Il pleut' writes down *the French word 'il'*."

[36] "The identification a word uttered or inscribed with one heard or read is not a matter of resemblance between the two physical embodiments.... Rather it is...a matter of intention . We depend heavily on resemblance between utterances and inscriptions...in order to divine these critical intentions.... We also take account of accent and idiolect and all the use clues to intention It is the latter that decides the matter." Kaplan, *op. cit.*, p. 104.

Of course, a performance may be in some sense a bad performance of a word. But, crucially, it is never a *non*performance: good or bad, it is a performance of that word. As Kaplan says:

> He may not do it well, from the external point of view. But it *is* what he is doing....No matter how poor the subject's *imitative* ability...we can imagine circumstances in which we would say, "Yes, he is repeating that name; he is saying it in the best way that he can." (103–04)

> ...the difference in sound or shape or spelling, can be just about as great as you would like it to be. (101)

Neither the emphasis on repetition (as a sufficient and pervasive condition of word matching) nor on the associated Constitutive Role of Intention is well motivated in Kaplan's discussion. Some straightforward observations will suffice to clarify our reservations.

First, notice that it is a grotesque exaggeration to suppose that when we produce a word we typically have in mind any particular performances which we intend to repeat. Intuitively, it seems that what we do is *pick up* a word and then, later, engage in performances with the intention of using *that word*. The original performances that allowed us to pick up that word are long gone from our mind or confused with others. The point is hardly unique to words. Introduced to a song or a dance, we may produce either without an episodic memory of the performances that allowed us to pick them up. To remedy this defect, an intermediate stage in word performance needs more emphasis.

Suppose we encounter some performance(s) that introduces us to a word. We thereby form an idea *W* of the word. This idea might outlive any memory of the performances that initiated it. We then engage in a performance that is controlled by an intention that involves *W*. At this juncture, it would be manifestly incorrect to describe the intention behind our performance as an intention to repeat, since there is nothing we are trying to repeat, nor anyone we are trying to mimic.[37]

To reinforce this picture, reflect on the myriad ways in which we can be introduced to a word. Although we might be introduced to it by a performance, we might just as well be introduced to it by description: "There is a word in English spelled 'b' followed by 'a' followed by 'b' followed by 'y' that means the same as 'infant'." This allows us to pick up the word and form an idea of it. Yet when we subsequently utter 'baby' we are obviously not trying to repeat a performance of it, since we were never exposed to any performance of it.

---

[37] In discussion it became clear that Kaplan never intended to subscribe to Pervasiveness as we describe it. Given this, it was certainly a strategic error for Kaplan to emphasize the importance of repetition At any rate, the idea presented in the above paragraph renders the concept of repetition altogether peripheral: when we speak we are certainly not repeating ideas.

Our first concern, then, is that (even noncreative) performances of a word rarely are controlled by a repetitive intention, and so Pervasiveness is *in*correct.

Our second worry concerns the presumed Constitutive Role of Intention. To begin, reflect on songs. In this case too, we can ask whether a performance being controlled by an intention to perform a particular song *S* guarantees that it is a performance of *S*. We take it to be obvious that in the case of songs a certain level of *tolerance* is in play: we allow a performance to be of a song even though it is a *bad* performance of it. But there also are myriad cases where even though there is a controlling intention to perform a particular song *S*, the performance counts as a *non*performance rather than merely a bad one: *tolerance has limits*. (It goes without saying that the boundary is vague, but that of course is no argument against the existence of the distinction.)

Suppose one intends to perform the song "Hey Jude" but merely grunts because of a defect or pathology: then the performance is a *non*performance, not merely a bad one. Suppose you intend to perform "Hey Jude" but mistakenly identify its lyrics and music with those of "Let It Be." Then your performance is a *non*performance rather than merely a *bad* performance of "Hey Jude" (leaving open whether it is a performance of "Let It Be").

The analogy between words and songs is arguably not perfect. In the case of songs, there is a sharp disanalogy between the relation of the vocal rendition of the song, on the one hand, and what occurs on the written score sheet for the song, on the other. The former is a performance of the song, whereas the latter is its mere representation. Arguably there is no such contrast between the spoken and written word (*contra* Aristotle).[36] But we do think that a nonperformance/bad performance distinction of the sort we just sketched carries over to words.

---

[36] We only say 'arguably': one line to pursue is that there is a deep analogy even here. Note that certain Kaplanian ideas about words carry immediately over to songs: (i) whether a performance counts as a performance of a song may be constitutively related to the community to which a performer belongs. Suppose a Martian wrote a symphony that resembles, but is still markedly different from, Beethoven's Fifth. Intuitively, a human performance may count as a bad performance of Beethoven's Fifth rather than a perfect performance of the Martian symphony owing to certain sociohistorical facts. (ii) Just as performances of a word may evolve in such a way as to change their character, so too with a song. As a song enjoys different renditions across time, its performances may change their character significantly. We return to these analogies in section IV.

If you intend to say the word 'dog' but instead merely grunt, then that is a nonperformance of the word rather than a bad performance. If Spooner intends to say, "The Lord is a loving shepherd," and instead says, "The Lord is a shoving leopard," then that is a nonperformance rather than a bad performance of what he intended to say. It is not that Spooner succeeds in his intention to say the word 'loving' but says it defectively (as when, for example, one mumbles or stutters); it is rather that he does not say the word 'loving' at all. If you see the word 'dog' for the first time and are ordered to rewrite it but proceed to write 'g' followed by 'o' followed by 'd', then you have failed to write the word you were asked to: your intentions, as a matter of fact, lack constitutive authority. In effect, we are advocating *Tolerance*:

> *Tolerance*: Performance p is of a word w only if it meets relevant performance standards.[39]

Tolerance is anathema to Kaplan presumably because he thinks it bucks the Evolutionary Constraint. After all, evolutionary considerations allow that a word w pronounced 'dog' in one epoch is pronounced 'god' in a later one. Isn't this evolutionary insight incompatible with our claim that you have failed to say the word 'dog' if you produce something with 'g' followed by an 'o' followed by a 'd'? We presume not, so long as care is taken to realize that whether a performance counts as tolerable does not supervene on its intrinsic features, but is instead determined by certain relational facts. Let us explain.

Assume, with Kaplan, that a word may evolve in its performance profile. Further suppose that the limits of toleration evolve as well: the line between nonperformance and bad performance itself may shift over time. Thus, for example, it is possible that there is a word w and two communities C_1 and C_2 such that the string 'god' is a tolerable way of writing down w in C_1, but not in C_2. These evolutionary considerations block sweeping generalizations such as "Nothing that looks like 'god' will ever count as a performance of the word 'dog'."[40] But this is perfectly compatible with the claim that if someone *in our community right now* writes down 'g' followed by 'o' followed by 'd', he will *not* have written the word 'dog' (even though he may have intended to do so).

---

[39] Standards are typically generous. There are cases where the community can be brought to see handwriting as inscribing certain words even though this is not initially apparent. Similarly, there are cases where a community can be brought to see that someone is performing a certain sonata even though this is not evident, given that the performance is so bad. But still, tolerance has its limits. If someone, owing to a visual problem, sees a 'B' as a 'P', intends to repeat someone who says 'Big', and writes 'Pig', then he fails to repeat the word he intended to repeat. In a footnote, Kaplan recognizes exceptions to the constitutive authority of intention, namely, cases where "*to the astonishment of the speaker* the wrong word came out . Some dark force has reached into the speaker's psyche and misdirected the hand of intention." Kaplan, *op cit.*, p. 105n11. But notice that the 'Pig'/'Big' case is not quite like that. The speaker in Kaplan's case reacts with astonishment just by virtue of having witnessed his or her own performance. Not so in the 'Pig'/'Big' case.

To accommodate the evolutionary insight, we need to acknowledge that the limits of toleration are *local*. What is a tolerable articulation of 'dog' will be local to a particular community. Sometimes we are more tolerant of speech defects, foreign accents, and unusual scrawls than at other times. Shifts happen. In general, a necessary condition on articulating a word is that the articulation passes the standards of the local community.[41] We call this *Tolerance**.

> *Tolerance**: Performance p is of a word w only if p meets relevant *local* performance standards on w.

In this way, the evolutionary insight can be harmonized with the thought that there are limits on toleration of the sort we have gestured at. Since this harmony can be achieved, we see little advantage to the more radical tack that Kaplan is proposing.[42]

In sum, Kaplan's vision could be improved upon in two ways. First, he should have allocated a less central role to repetition. This need not mean we should ignore historical aspects of words. It is rather merely to allocate a suitably central role to the mediating ideas of words—ideas that may outlive memories of particular performances. Moreover, even if the object of those ideas is socially determined, it is going too far to *require* that we must encounter performances of a word in a community as a precondition of forming an idea of that word. The second way in which Kaplan's vision can be improved upon is to recognize that he overstates the constitutive importance of intentions. In order for a performance to count as a performance of a word, it needs to fall within locally acceptable limits. Not anything goes.

---

[40] Furthermore, not just anything that looks like 'dog' counts as a performance of the word 'dog'. In this regard, though we agree with Herman Cappelen that intentions play a lesser role than Kaplan affords them, we disagree with him about whether the relationship between shape and expression is context insensitive. See Cappelen, "Intentions in Words," *Noûs*, xxxiii, 1 (1999): 92–102.

[41] It is more accurate to require that one passes the standards of the relevant community, where in general, but not always, the relevant community is local. If someone is doing a performance in an attempt to simulate a fourteenth-century English speaker, then the relevant community will *not* be the physically local one. And of course there need not be pressure on an American speaker to produce the word 'schedule' in the British manner even if he is sitting at High Table. Indeed, it is preferable that he *not* try to mimic the local customs

[42] It is also obvious that the tolerance conditions will be articulation-system dependent. Within a particular community there will be conditions on whether a vocal performance counts as an articulation of the word 'dog', and rather different conditions on whether a written or signed performance counts as one.

III. COMMON CURRENCY NAMES

The relevant sense of 'word' that forms Kaplan's putative focus—one that he takes to satisfy all three constraints from the last section and to be urged by them—is supposed to be one according to which two naming practices involving the string 'Peter' turn out to articulate two different names—that is, two different common currency names. There is an obvious sense in which a single name is involved here. In deference to this sense, Kaplan claims that there is a single 'generic name' in play. But he also claims that in such a case there will be a proliferation of 'common currency names', and in particular there will be a distinct common currency name associated with each distinct naming practice that uses 'Peter'. Indeed, it becomes clear that an important motivation for Kaplan's discussion of words is the demarcation and vindication of the category of common currency names.

For our part, we do think that there is one or more important sense of 'word' that answers to the Evolutionary, Multiplicity, and Coincidence constraints. But we harbor considerable skepticism about the category of common currency names. It is the project of this section to expand upon and justify that skepticism.

Kaplan makes it clear that he does not intend common currency names to be the designata of quote names—say, 'Peter'—as used in English, or at least as in philosopher's English. Certainly, such a theory would make a hash of the data, including the fact that it is perfectly acceptable to say, "They are both called 'Peter'." Rather, his idea is that there is a natural kind worthy of study which is important to philosophy and linguistics, and which answers to a way of theorizing about words that is prevalent in foundational philosophy of language and linguistics. According to this way of theorizing, various utterances, for example, of 'Paderewski' by English speakers (recalling Kripke's famous example—see Motivation Five below), may not all be utterances of the same name.[43]

The relation between the ordinary conception of names and the theoretical conception of common currency names is delicate. We mention three options.

First, one might hold that in typical contexts the extension of 'word' in ordinary language is the set that includes common currency names, explaining away contrary data by claiming that ordinary people are prone to radical errors in judgments about words. For example, when ordinary people say things of the form "Two of us have the same name," or "That name can be used to refer to several different people," they typically express falsehoods. We see this position as something like a last resort. (To be clear, we see no evidence that Kaplan holds such a position.)

---

[43] See Mark Richard, *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them* (New York Cambridge, 1990); Richard K. Larson and Peter Ludlow, "Interpreted Logical Forms," *Synthese*, xcv, 3 (June 1993): 305–55, Fiengo and May, *op. cit.*

Second, one might hold that the extension of 'word' in ordinary use is never, or is hardly ever, a set that includes common currency names. On this conception, common currency names are irrelevant to ordinary 'word' talk. So, for example, even when people say "You and I have the same name," they may be speaking the truth because they are *not* expressing a proposition that requires a shared common currency name for its truth.

Certainly, Kaplan allows that in many contexts we are not talking about common currency words when we deploy the word 'word'. As noted above, Kaplan admits the notion of a generic name and allows that two people can share the same generic name. It is clear, in context, that he is thinking that when people say "We have the same name," they may be expressing the truth that people share the same generic name.

Yet it is hard to construe him as merely positing a class of entities—common currency names—that is theoretically important for linguistics, philosophy of language, and cognitive science, but of which ordinary folk have no conception. After all, much of what he says fails to square with this. For example, he frequently speaks of people intending to repeat a common currency name. But how can someone intend to repeat a common currency name if she has no ability to have singular thoughts about particular common currency names and has no concepts whose extension is the set of common currency names? There is nothing untoward about introducing a class of theoretical entities beyond people's ken. But it would be very odd at the same time to accord ordinary intentions towards those entities a crucial role in one's account of their place in the world.

On a third approach, ordinary people are *sensitive* to the presence of common currency names, and in some contexts they use 'word' to pick out a class that includes common currency names.[44,45] On this

---

[44] Our taxonomy is not exhaustive. For example, here is a fourth position which we find less appealing. There are such entities as common currency words, ordinary people are aware of their existence and think about them, but ordinary people never use 'word' to talk about them.

[45] Note, at the other pole, that there are contexts in which we use 'word' in a way that excludes both common currency names *and* generic names. Thus, for example,

approach we are at some level already accustomed to distinguishing different common currency names even if there are plenty of contexts in which our use of 'word' is insensitive to such distinctions.

On the first version of the project, one would expect there to be ordinary judgments about the sameness or difference of common currency names but should be very skeptical about their reliability. On the second version, one would expect there to be no ordinary intuitive judgments as to whether common currency names are the same or different. On the third version, such expectations would not obviously be out of place; nor would one expect them to be obviously unreliable. Kaplan's project is clearly of the third sort. His pronouncements make little sense otherwise. We shall continue the pursuit of common currency words in that spirit.

As we have already noted, it is central to Kaplan's vision that different common currency names may be vocalized by the string 'Peter' when those vocalizations belong to two different naming practices. Thus, for example, we use a different common currency first name for John Donne than we do for John Travolta, even though each name gets articulated by similar-sounding performances. (That is not to say there might not be cases where we confuse these two individuals, and thus, where there is a single common currency name that we attempt to use both to refer to John Donne and to John Travolta. Suppose, for example, many different individuals were involved in the so-called Jack the Ripper crimes.) In addition, as already noted, Kaplan recognizes generic names, where John Donne and John Travolta, while they have different common currency first names, share the same generic first name. But if we take ordinary language data at face value, we automatically ought to be suspicious of common currency first names. We invariably speak of John Donne and John Travolta as sharing a first name, and it is rather difficult to access a reading of "John Donne and John Travolta have different first names" under which it comes out true. Something needs to be done, therefore, to justify positing common currency names as corresponding to one important conception of names. Something needs to be done to motivate Kaplan's claim that "for serious semantics...it is my common currency conception that [is] important" (111). In what follows, we canvass and evaluate five such motivations.

---

in contexts where we ask how many words someone knows, both generic names and common currency names get excluded. Perhaps in part influenced by such ordinary usage, one or two linguists whom we have encountered claim that names are not words at all. However, most linguists we have encountered doubt that there is any deep insight encoded by such remarks.

Motivation One: Words and Mentalese. Suppose, as many presume, that thought is conducted in an internal language—an 'I-language' or 'Mentalese'. It is eminently natural to hold that insofar as we distinguish two people in thought, we employ two different Mentalese tags for them. (One motivating thought is that we could not properly keep track of the sameness and difference of our own thoughts if there was pervasive semantic context-dependence even at the level of Mentalese.) Given this eminently natural picture, shouldn't we say that the public language performances of their names are articulations of two different names even if they sound alike?

Reply: This motivation plays illicitly on the verb 'articulate'. Suppose someone sometimes speaks French and sometimes speaks English. There may be some Mentalese expression for *rain* that he sometimes articulates using 'rains' and sometimes articulates using 'pleut'. But that does not mean that 'rains' and 'pleut' are the same word. There can be distinct words corresponding to the same Mentalese word.

Meanwhile, one might have different Mentalese vehicles in play when one uses 'red' to mean that something is red on the outside and when one uses it to mean that something is red throughout. But this hardly means two English words are in play. Likewise, a Spanish-to-English translator may "articulate" one Spanish word by different English words depending on context. But that hardly means there are many Spanish words being translated.[46]

Motivation Two: The Intuitiveness of Kaplan's Three Constraints. Does the fact that there is something very intuitive about the ideas behind the Evolutionary, Multiplicity, and Coincidence Constraints establish that we already have a pre-theoretic conception of common currency names?

Reply: As Kaplan himself acknowledges, *even generic names share this tri-fold profile* (115–16). The generic name 'John' was pronounced very differently in Shakespearean times than it is today; it also can be realized in many different systems of articulation—it can be spoken, written, signed, Brailled, semaphored, and so on; the names 'Shawn' and 'Sean' are pronounced the same, though there are obviously contexts in which one ordinarily can count them as two names. In short, the intuitiveness of Kaplan's three ideas need not reveal any tacit commitment to an ontology of common currency names nor, in particular, to the thesis that in some ordinary sense of 'word' there are many different words that English speakers express by 'John', individuated by naming practices.

---

[46] It bears emphasizing that while an individual might have two Mentalese words MW_1 and MW_2 for a particular pair of individuals *X* and *Y*, which he expresses using the same 'generic word' of English, he may *not* have two Mentalese names for those Mentalese words. Hence, he may not be capable of intending to use one of those Mentalese names rather than another. Thus, supposing MW_3 is the Mentalese name for the generic name 'John', he may be able to have the thought that the (generic) name 'John' is sometimes used as a name for *X*, but not that the Mentalese name MW_1 is a name for *Y* On a related point, while he may be able to intend to use 'John' as a name for *X* on a particular occasion, he may be unable to intend to use 'John' to translate MW_1 on a particular occasion We have already seen that Kaplan needs common currency words to figure in the intentions of language users. Even assuming multiplicity of Mentalese names, they may be unsuitable to ground the presumed multiplicity of word-theoretic intentions.

Motivation Three: Common Currency Names Are Needed for Kripkean Theories of Reference. Consider the following, fairly standard, simplified Kripkean model of reference:[47]

Stage One: A speaker associates a referent with a name by either a reference-fixing description ("Let 'Johnny' name the tallest male model in Alaska.") or by ostensive baptism ("We christen this ship 'Johnny'.").

Stage Two: That name is passed through the community by a causal chain of communication (transference). The reference of later uses of that name is determined by a historical chain that leads back to the reference-introducing event. The fundamental rule is this: the reference of a name (insofar as it has a reference) is determined by the Stage One event that associates some object with that name. Meanwhile, an appropriate causal chain renders some performance by an individual a performance of the same name that was introduced by a Stage One event. In short, a performance p_2 is of a name n only if there is an appropriate historical-causal chain connecting p_2 and some reference-introducing performance p_1 of n.

As Kaplan is himself aware, this model requires a conception of names that aligns with the common currency conception (93–95).[48,49] There is a many-to-one relation between generic names and Stage One events. If the same generic name 'John' is associated with John Donne and John Travolta, then it makes no sense to speak of the event whereby the reference of the generic name 'John' was fixed.

---

[47] Saul Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980); Michael Devitt, *Designation* (New York Columbia, 1981), and Richard, *op. cit*

[48] Richard writes, "Part of what makes for sameness of name in the requisite sense is being part of the same 'causal chain' of transmission." *Ibid.*, p. 183.

[49] Kripke (*op. cit.*, note 9) writes, " .two totally distinct 'historical chains' that by sheer accident assign phonetically the same name to the same man should probably count as creating distinct names despite the identity of the referents." His commitments on the present topic turn on how 'totally distinct' is to be read. Is the use of 'John' to talk about John Donne '*totally* distinct' from the use of 'John' to talk about John Travolta?

For on the generic conception, 'John' can be used to refer to a variety of different people depending on context.

Reply: Granted, one cannot take on board the simplified Kripkean model in all its aspects and simultaneously hope that generic names will play the role allotted to names within the model. Yet one wonders how damaging it would be to alter the Kripkean model slightly, so as to eliminate the need for an ontology of common currency names. By our lights, what is crucial to the 'chain' idea is that the reference of particular *uses* of a name is typically deferentially determined. Suppose, for example, you encounter a particular use of the generic name 'John' by some utterer *U*. You may then use that generic name where your controlling intention is that your *use* of 'John' refer to whichever object *U* was referring to with his use of 'John'. In this way, the reference of the generic name 'John', on your occasion of use, is inherited from the reference of that name on *U*'s occasion of use. The central inheritance idea is thereby preserved *without* relying on an identity of a common currency name.[50]

Notice that this kind of model (unlike Kaplan's Kripkean one) can also readily account for cases where you quite obviously *inherit* reference without using the same generic name. Thus, suppose we hear someone talking about a guy called 'John', and he attributes some nasty acts and traits to him. One of us may decide to start calling him 'Mr. Nasty'. Here the reference has nothing to do with the preservation of a name. What is crucial is that the controlling intention is to use 'Mr. Nasty' in the relevant context of conversation to refer to whomever 'John' had referred.

In short, while one Kripkean theory of reference requires serious reliance upon an ontology of common currency names, it is not clear that any important insights would be lost by moving to a variant model that eschews any such reliance.

Motivation Four: Ambiguity and Polysemy. Linguists standardly distinguish ambiguity—as exemplified by the unrelated meanings for 'bank'—from polysemy—as exemplified by the family of related meanings for 'healthy' (as applied to animate objects, animal excretions, and food) and 'window' (as applied to both the physical object and its frame, as in "He opened the window" and "He went through the window"). In the case of ambiguity, we would not be surprised to find two utterly different words in other communities, but in the case of polysemy we would. A standard piece of linguistic lore is that with ambiguity two words are in play but with polysemy only one with a variety of connected semantic potentials. In order not to prejudge certain questions, we will employ the technical term 'lexeme' for posits of linguist orthodoxy.

---

[50] There are slightly more complex models. For example, one might forge a Mentalese tag *T* for the object picked out by some use of a name *N*, and then use *T* to specify the intended referent of some later use of *N* (Our wariness of the pervasiveness of repetition may encourage moving in the direction of some such model.) This model allows that some name *N* may on varying occasions have its reference fixed by differing Mentalese tags. But to infer that this implies that differing names are used on those occasions would be to commit the fallacy identified under Motivation One

Various considerations induce linguists (and philosophers[51]) to say two lexemes are associated with 'bank'. One concerns distinct etymologies. It generally is taken to be a condition of ambiguity that the lexemes in question are known to have developed from what were formally distinct lexemes at some earlier stage: the use of 'bank' that means *a raised shelf or ridge of ground* has its roots in the Old Norse word 'banke', and its use that means *financial institution* has its roots in the Old French word 'banc'. Another consideration includes distinct distributional reflexes. For example, by and large, anaphora and ellipsis work differently for ambiguity and polysemy. We say, "He opened the window and went through *it*" (where 'it' is anaphorically tied to 'window' even though the two occurrences have slightly different meanings—physical object versus frame); similarly, we say, "He is healthy and *so is* the food he prepares for his family," but we do not say (except as a pun), *"He put some money in a bank and then swam to *one*," or *"After losing forty pounds, he is light and *so is* the color of his hair."[52]

Positing lexemes offers a potentially attractive theoretical explanation of these contrasts: an anaphor can be tied to the original only if the lexeme in play in the original is appropriate to the environment in which the anaphor appears.[53]

Mightn't we appeal to lexemes in vindicating this aspect of Kaplan's perspective?

---

[51] "If one of the tokens [of 'bank'] refers to a financial institution and the other to the edge of a river, it is implausible to insist that they belong to the same type." Zoltán Gendler Szabó, "Expressions and Their Representations," *The Philosophical Quarterly*, XLIX, 195 (April 1999): 145–63, at p. 148.

[52] As a heuristic, linguists sometimes appeal to dictionary entries. While one expects multiple distinct dictionary entries for 'bank', one would not for 'red', 'window', or 'healthy'.

[53] It is not clear that the historical and distributional ideas are really complementary. Consider 'right'. It is standardly treated as ambiguous—largely for distributional reasons. But the uses of 'right' to mean *a direction* and to mean *correctness* have a common etymological origin. In the end, historical evidence plays an evidential, not a constitutive role in the thinking of most contemporary linguists on the subject: if 'right' has two distinct 'analyses' flowing from distinct 'lexical entries', then there is more than one lexeme in play even if there is a unity in historical origin for both analyses.

Reply: We have already acknowledged that there is flexibility in which objects we talk about in using 'word'.[54] Sometimes, for example, we treat quotation words as picking out something that can be either written or spoken, but other times as picking out the very sign used to articulate words. Thus, for example, we might in one context agree that 'red' is the same word as 'RED' (using a more abstract sense of 'word'), while in another context agree (using quote words to name the signs themselves) that 'red' is in lower case, while 'RED' is not. (Consider similarly, "'red' occurs three times in the second paragraph, but 'RED' only once.")[55]

We concede further that there are uses of 'word' that seem to point in the direction of lexemes. The relevant frame of mind here is one that Lyons invoked in his seminal discussion of lexemes,[56] one where we are willing to make such claims as:

(1) 'Find' and 'found' are versions of the same word.

Claims like (1) abstract away from the surface in a quite radical way. In that frame of mind, it is natural to say (2):

(2) 'Will' and 'willed' are versions of the same word.

In that frame of mind, it is also natural to say (3):

(3) 'Will' and 'would' are versions of the same word.

But it is not natural to say that 'willed' and 'would' are versions of the same word. In this frame of mind, then, it would seem that we are committed to there being *two* words associated with 'will'. Let us grant, then, that there are contexts in which 'word' is being used to pick out lexemes.

Crucially, however, it is very far from clear that this can serve as a basis for vindicating Kaplanian common currency names. To do so would require that we say that two different lexemes are manifested by 'John' as it occurs in 'John Donne' and 'John Travolta'. But what justifies this? The fact that different people are denoted in the various contexts of use does not settle the issue—after all, polysemous uses of a single lexeme generate varying denotations as well. Insofar as the data provide guidance, they do not seem to support an ambiguity thesis—after all, it is felicitous to say, "She is Janet and [pointing to a different person] so is she," and "Every Janet that I know...." The felicity of such statements tells against ambiguity and hence lexeme proliferation.[57] To claim that there are many lexemes associated with the use of 'John' in English is simply bad linguistics. Insofar as names are identified with lexemes, that tells against Kaplan's common currency views, *not* in favor of them.

---

[54] In the language just introduced, perhaps the right conclusion is that 'word' is itself polysemous.

[55] For an explanation of the phenomena that does not invoke ambiguity or context sensitivity, compare with Cappelen and Ernest Lepore, *Language Turned on Itself. The Semantics and Pragmatics of Metalinguistic Discourse* (New York: Oxford, 2007).

[56] Lyons, *Semantics. Volume 2*

Motivation Five: Attitude Reports. Some accounts of attitude reports rely heavily on lexical considerations, treating some attributions as true only if the speaker bears a certain relationship to words in a report's complement clause.[58] In these contexts, it is alleged that form trumps content and reference. (4) and (5) attribute distinct beliefs to Lois because they relate her to different names.

(4) Lois Lane believes that Superman can fly.
(5) Lois Lane believes that Clark Kent can fly.

Insofar as one goes in for this type of account, there is pressure towards admitting that a common currency conception of names is central to ordinary practices of belief reporting. After all, this account is supposed to extend to Kripke's well-known 'Paderewski' case.[59]

Peter is given two bits of testimony from two different people: "Paderewski has musical talent," and "Paderewski does not have musical talent." Peter does not realize (let us suppose for a good reason) that each piece of testimony is about the same person and so accepts both pieces of testimony. We look on. We are tempted to say, "Peter believes that Paderewski has musical talent," "Peter believes that Paderewski has no musical talent," and that Peter is not being irrational. The solution proposed by the relevant accounts is that the pair of reports relates Peter to distinct names. But this apparently requires that there are two different names in play and that this fact is tacitly recognized by our reporting practices. Thus,

---

[57] We again note in passing that some contemporary linguists refuse to treat names as words, but we put that issue to one side.

[58] For example, Gilbert Harman, "Logical Form," *Foundations of Language*, IX, 1 (September 1972): 38–65; James Higginbotham, "Linguistic Theory and Davidson's Program in Semantics," in Lepore, ed., *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson* (New York: Blackwell, 1986), pp. 29–48; Higginbotham, "Belief and Logical Form," *Mind and Language*, VI, 4 (Winter 1991): 344–69, Kaplan, *op. cit.*; Larson and Ludlow, *op. cit.*; Larson and Segal, *Knowledge of Meaning* (Cambridge: MIT, 1995); Richard, *op. cit.*; and Segal, "A Preference for Sense and Reference," this JOURNAL, LXXXVI, 2 (February 1989): 73–89.

[59] Kripke, "A Puzzle about Belief," in Avishai Margalit, ed., *Meaning and Use* (Boston: D. Reidel, 1979), pp. 239–83.

Ludlow and Larson infer that these homophonous expressions "are actually two names here, 'Paderewski$_I$' and 'Paderewski$_{II}$'."[60]

Reply: Note first that if this style of account is right for identity confusion, then one ought to expect it to apply to a range of cases where proper names are not in play. For example, suppose someone is told, "No one has ever lived here," and then later told, "Someone once lived here," and while he has excellent evidence that 'here' refers to different places, in fact it refers to the same place twice over. This has a very similar structure to the Paderewski case, and surely demands a similar solution. But the solution in question, as applied to this case, requires us to say that two words are in play, 'here$_1$' and 'here$_2$', and not one. Certainly, some will be unwilling to absorb this consequence—but it is important to see that it needs to be absorbed.

Our main objection to the account in question is that it flies in the face of something that is right about Kaplan's account. We do not think merely intending to use a word one has picked up guarantees success. However, if one satisfies local performance standards, it is hard to deny that intending to use a word will bring success in its wake. Now, suppose someone at one time tells us "Paderewski is musical," and at another time, "Paderewski is not musical." We repeat back, "Paderewski is musical," and "Paderewski is not musical."

---

[60] See Ludlow and Larson, *op. cit.*, p. 319. A related but possibly distinct motivation focuses on logical truth itself. We say 'Hesperus is Hesperus' is a logical truth, but not 'Hesperus is Phosphorus'. This differential judgment appears to rely on the recognition that 'Hesperus' and 'Phosphorus' are distinct names. But suppose now that there are two different naming practices wherein the generic name 'John' is associated with an individual. (Suppose John Lennon and John Travolta are in reality one object with two personae. Isn't it tempting to think that 'John is John' is not a logical truth even though the same generic name appears on each side? If name individuation is to provide the basis for an account of logical truth, it seems that generic names cannot do the work. Considerations such as these are evidently important to Kaplan's own motivations. See, for example, Kaplan, *op. cit.*, p. 94. Now, Kaplan is sensitive to the fact that even common currency name identity in '$a = a$' will not secure a sense of triviality in every case given that one might store the same common currency name twice over without realizing it. But one might hold that in such cases there is a logical truth that the speaker fails to recognize. By our lights, an appeal to common currency names will only be appealing at the foundations of logic if one is willing to make similar appeals to a fine-grained ontology of the words 'here' and 'that', appeals about which we are skeptical. (The point extends to predicates: for any context-dependent predicate, F, a vindication of the claim that 'All Fs are F' is true by virtue of its form will have to rely on a fine-grained ontology of predicates.) Further, we note that given that common currency names arguably can switch reference, it is a short step to allowing that there are cases where some uses of a single common currency name refer to one individual, others to another, and hence, cases where an identity claim flanked by two occurrences of the same common currency name is false. The fact that even common currency names arguably can have variable referents entails that they cannot, after all, very well play the role of the individual constants of logic. This all raises important issues about the form-theoretic conception of logical truth that we cannot pursue further here.

Suppose we have good evidence she is speaking about different people but is trying to confuse us. It is clear enough that the speaker is using the same word in both performances. It is also clear enough that on each occasion we intend to repeat the name we are confronted with and that we satisfy local performance standards for that name. Given all this, it is hard to deny that we use the same word as the speaker on each occasion of repetition. Given symmetry and transitivity of identity, it follows that we both use the same name on each occasion. But this is hard to square with an account of the relevant belief ascription that depends upon our using *different* names on each occasion.

We do acknowledge that something in the vicinity of these accounts nevertheless may be correct. Ludlow offers an account of the Paderewski case according to which we, the ascribers, make up two names in some internal language on the fly with which we label the Paderewski nodes of the semantic trees. Our sense that we are not convicting the subject of incoherence is then explained by the fact that we use different labels in each of the ascriptions.[61] Here is not the place to evaluate this account, which relies on labeling opportunism. But note, crucially, that this account does not require that the subject of the belief ascription use two different names in a Paderewski case. We suspect that insofar as there is anything to the idea that Paderewski cases are to be explained by associating the subject with two different labels, it will not rely on an ontology of common currency names.

In sum, we have yet to see a powerful case for the thesis that in some ordinary contexts, 'word' has Kaplanian common currency names in its extension. Perhaps a fine-grained ontology that associates multiple lexemes with 'bank' can be justified on theoretical grounds, and perhaps there are settings in which we use the word 'word' to talk about lexemes. But even so, the case remains to be made that Kaplanian common currency names are the objects of ordinary repetitive intentions or ordinary 'word' talk. Certainly, none of the justifications we have examined are particularly persuasive.

IV. CRITERIA OF IDENTITY

We have voiced multiple reservations about Kaplan's account, and along the way sketched an answer to the second of the pair of questions with which we began. But what of the first question?

Criteria of identity come cheap if we allow them to go trivial. To what extent can we hope to come by both true and informative criteria of the identity of words? Given the flexibility in our uses of 'word', the context of our inquiry needs to be made clearer before we press ahead. The preceding discussion indicates that the theoretically deepest—and of most interest to philosophers—notion of 'word' is one according to which it is used in a way that conforms to the lexeme conception. We ask the reader to understand us to be using 'word' in that way, and to read what follows in that light. We shall suggest that even so restricted an expectation of identity criteria for lexemes to a large extent is misplaced.[62]

---

[61] Ludlow, "Interpreted Logical Forms, Belief Attribution, and the Dynamic Lexicon," in Katarzyna Jaszczolt, ed., *The Pragmatics of Propositional Attitude Reports* (New York: Elsevier Science, 2000), pp. 31–42

On the lexeme construal, we must institute a sharp distinction between words and their performances. With that distinction in place, pursuit of a criterion of identity can take either of two forms. (Here we borrow some terminology from Williamson.[63])

First, we may attempt to articulate interesting *first-level identity criteria*, where the objects for which the criterion of identity is stated are the same as those between which the criterial relation obtains. The criterion of identity for sets stated by the Axiom of Extensionality (sets are identical iff they have the same members) and Davidson's criterion of event identity (events are the same iff they have the same causes and effects) are examples of first-level identity criteria.

Second, we might pursue a *second-level identity criterion*, where we state a criterion for two objects from a different domain to stand in some particular relation of interest to the same object from our original domain of interest. The Fregean criteria of identity for directions and numbers are both second level. Two lines (neither of which is itself a direction) are of (here 'of' marks the particular relation of interest) the same direction iff they are parallel; two classes are numbered by the same number (here 'numbered by' marks the particular relation of interest) iff there is a one-one map from one to the other.[64]

---

[62] Certain considerations that follow will carry over to uses of 'word' that are slightly less abstract than the lexeme use. We shall not, however, be asking whether and to what extent there are identity criteria corresponding to each use of 'word'.

[63] Williamson, *Identity and Discrimination* (Cambridge. Blackwell, 1990).

[64] At either first- or second-level identity criteria, we might less ambitiously seek only (interesting) necessary or sufficient conditions. Thus, at the first level, we might seek interesting claims of the forms:

Words w_1 and w_2 are identical only if R holds between w_1 and w_2.
If R holds between words w_1 and w_2, then $w_1 = w_2$.

Meanwhile, at the second level we might look for interesting claims of the forms:

Performances p_1 and p_2 are of the same word w only if R holds between p_1 and p_2.
If R holds between performances p_1 and p_2, p_1 and p_2 are of the same word w.

Also, at both first- and second-level criteria of identity, we might pursue something more ambitious. Take the above second-level criterion for directions. It is an intra-world

Note that it should not strike anyone as particularly alarming if we fail to devise informative first-level criteria of identity. After all, the standard criteria of identity for directions, for example, are second level and *not* first level. (The prospects for informative first-level criteria of identity for directions would be made worse if we allowed for directions that are not the direction of any line. Assuming there are unarticulated words, the situation with words is akin to that bleak situation.) Nevertheless, let us consider an attempt at a first-level criterion for word identity from Richard and Millikan:

> *Origin*: Words w_1 and w_2 are identical iff w_1 and w_2 have the same originating event (where the originating event is the first performance of a word).

A few observations about Origin are in order. First, it assumes that the relation 'performance of' between an event and a word is already (sufficiently) well understood. Second, it assumes there are no unperformed words. And third, it assumes there cannot be a tie as to which events are the first performances of a word. Origin's take-home message, in short, is that there is exactly one event that is the first performance of any word. (Origin is equivalent to that message, since the latter is obviously derivable from the former and, given Leibniz's law, the former is derivable from the latter.) As such, Origin is far less informative than one might initially think. For example, it is logically compatible with Origin that only one word exists: to derive the obviously correct conclusion that 'cat' is *not* the same word as the word 'dog' you need additional information about which was the first performance of each. In any case, if what we have said about unarticulated words is correct, at least one of the assumptions upon which this first-level criterion depends is *not* true.

The following first-level criterion *is* arguably correct:

> *Performance*: Words w_1 and w_2 are identical iff it is not possible that some performance is of w_1 but not of w_2.

Unfortunately, all that Performance encodes is that *if* there are two distinct words, then it is possible that there be a performance of one without being of the other, and this does *not* seem particularly interesting.

---

criterion; it tells us when two lines in the same world partake of the same direction. But it does not obviously tell us anything about when lines in different worlds partake of the same direction. One might not think that *being parallel to* is a relation that can obtain between things in different worlds.

Do second-level criteria of identity fare any better? In the case of directions, we have an adequate second-level criterion already in hand: lines share the same direction iff they are parallel. (Note that this second-level criterion would not be impugned even if there were directions that were not the direction of any line.) With words and their performances, the situation appears to be much less promising. We need to face the prospect that there is no easily statable relation between performances that provides an interesting condition that is also both necessary and sufficient for their being performances of the same word.

Let us begin by engaging with a more modest task, namely, that of providing interesting necessary *or* sufficient conditions.

A candidate sufficient condition, Kaplanian in spirit, is *Intention*:

> *Intention*: If performance p_1 is ***intended*** to be a performance of the same word w as performance p_2, then p_1 is a performance of the same word as p_2.

We have already seen that Intention fails: among other things, certain grunts and groans are not performances of words no matter how much we intend them to be so.

A candidate for an interesting necessary condition on word identity, also inspired by Kaplan, is *Connection*:

> *Connection*: If p_1 and p_2 are performances of the same word w, then there is a performance p_3 of w such that p_1 and p_2 are both ***historically/causally connected*** to p_3.

Connection also fails, and for reasons we have already provided: performances p_1 and p_2 might be built out of morphemes presented separately to each performer, with neither traceable back to a common cause that takes the form of a performance of the word w of which each is a performance.

We do ***not*** however wish to preclude necessary conditions that are philosophically interesting. For example, a much weaker version of the causal idea to which we are somewhat more sympathetic is *Isolation*:

> *Isolation*: If linguistic communities c_1 and c_2 are causally isolated, then performances p_1 in c_1 and p_2 in c_2 are not performances of the same word.

Some have the intuition that a perfect intrinsic match among performances suffices for a word match.[68] But this intuition dissolves once one sees that the performance profile of a word can evolve.[66] Suppose two isolated communities have words pronounced as 'moan' and 'mean', respectively. There is no inclination to say that a single word is in play. Suppose the performance profile of 'mean' evolves so that its performances are 'moan'-like. We are tolerant enough evolutionarily to allow that the same word is in play in the latter and earlier performances. But the logic of identity forces us to admit that different words are in play in latter performances of 'mean' and current 'moan' performances of the other community, despite an intrinsic match. (Of course, it is obvious that no second-level criterion of identity can be wrought from Isolation. At best we get a weak, though mildly interesting, necessary condition for word matching.)

---

[68] See, for example, Cappelen, *op. cit.*, pp. 95–96.

The task of providing an interesting necessary condition on word matching between performances did not prove hopeless. And our survey of candidate sufficient conditions has been far from exhaustive. Ought we to be pessimistic about the possibility of a satisfying second-level criterion?

In this connection, it is worth underscoring a particularly satisfying feature of the second-level criterion of directions: namely, the relation of being parallel is a relation we can grasp independently of an ontology of directions. Call a second-level criterion 'quasi-reductive' if the relation that appears on the right-hand side of the criterion is capable of being grasped independently of the domain for which one is providing the second-level criterion. In the case of words, the hope for a quasi-reductive criterion seems rather dim. Any criterion that makes reference to intentions to repeat a word, to use the same word, and so on will be inappropriate to the quasi-reductive project. And we see no prima facie reason for optimism that the stock of relations compatible with the quasi-reductive project can provide the basis for a criterion that comes close to being compelling.[67] In short, the relations that are naturally appealed to when concocting a second-level criterion are relations that cannot underwrite a quasi-reductive criterion.

Of course, assuming both the existence of words and that the facts about word matching as between performances supervene on the microphysical ground floor, there presumably will be some extremely complicated second-level criterion that is extensionally correct, and some yet more complicated second-level criterion that will be necessarily correct. However, this by no means implies that we have epistemic access to such criteria, nor that the success and well-functioning of thought and talk about words relies on some tacit grasp of such criteria.

---

[66] It also dissolves once one gets into the lexeme-theoretic conception of words, according to which 'will' and 'would' are versions of the same word, 'will' and 'willed' are versions of the same word, but 'willed' and 'would' are not versions of the same word.

[67] It goes without saying that the hunt for interesting inter-world criteria will be even more problematic.

Now, from the armchair we have no general proof that compelling second-level criteria are unavailable. Certainly, the abject failure of attempts to provide criteria of this sort does not inspire hope that an interesting second-level criterion will be forthcoming. But in advocating pessimism about second-level criteria we do not merely wish to rely on a pessimistic induction. We shall, in what follows, sketch a picture that we believe is independently plausible and which suggests that the pursuit of second-level criteria is rather misplaced. We shall, along the way, ask a related, urgent question: why is an ontology of words-as-lexemes legitimate, even in the absence of a second-level criterion of identity?

To begin, it is useful to recognize that the situation with words is in many ways analogous to the situation with dances and games. In each of these cases, the following features are in play. First, community members have a pretty good appreciation of locally acceptable performance standards. That is to say, they have a pretty good capacity to distinguish performances from nonperformances in their local environment. Second, community members have an evolutionary conception of the relevant entities: dances may have been performed differently and games may have been played differently in earlier epochs. But, third, there is vast indecision about the distinction between cases where a new dance or game comes into existence and cases where it is merely an old game or dance in different clothing.[68]

Notice that an appeal to some crude causal criterion is of little use in resolving such indecision. Cases where a new dance is inspired by an old dance and cases where old and new performances are varying performances of the same dance are all cases of causal connectedness between the old and new.

As with dances, there will be vast indecision as to when words comes into existence. Consider, for example, 'moan' and 'mean'. Most English speakers take themselves to know that local performances of each are performances of different words, and hence, they take themselves to know that 'moan' and 'mean' are different words. Yet when confronted with the historical facts we have no clear judgment as to when each sprang into life.[69] There are causal lines

[68] It is sometimes said that all that holds various games from different locales and epochs together is their name. Obviously, this maneuver, whatever its other merits, will not suffice for individuating words

[69] Cf Cappelen, *op. cit*, p. 95.

from both clusters of performances to a cluster of performances in Saxon England. We are less tempted to say that there were two words in play at that point. And yet we have no clear judgment as to when our words were first performed in the history of speaking peoples. Similarly, we have no clear judgment as to whether some distant causal predecessor of this or that word among Latin or Greek peoples was our word or merely an inspiration for it. This indecision will not disrupt actual practices very much—the discriminations we are called upon to make in practice are almost always local, and indecision in this domain is not nearly as rampant.

As we have already hinted, when an ontological posit has the kind of profile just gestured at, we can scarcely hope to defend its legitimacy on the model of the abstraction principles standardly invoked as foundational to the talk of directions and numbers. Even within our own locale, nothing cleanly marks word matching between performances in the way that parallelism marks direction matching between lines or how the existence of a one-one function marks number matching between pluralities. And the situation is worse still when we move beyond that locale, owing to the vast stretches of indecision.

It is common enough for the metaphysician to encounter the kind of profile just described. There are two standard reactions. The natural-kind optimist supposes there are natural contours in the world that favor certain candidates over others: while our knowledge of word matching may not favor one candidate (or one small cluster of candidates) over others, one may still count as the semantic value on account of its being more natural, joint-like, and less gerrymandered than the others. Hope might even be held out that empirical inquiry will uncover some such boundaries. Perhaps, for example, empirical inquiry about the structure of morphology or psycholinguistics will resolve questions of word matching one way rather than the other.

A second reaction—that of the gruesome pessimist—is the same as the first except that it does not assume underlying joints in nature that make it natural to resolve questions one way rather than another.[70] Instead, there is a plenitude of packages of objects and exemplification relations in the vicinity of talk about words, none more natural than the others, each fitting the extent of our knowledge about word matching. The candidates give different verdicts to questions about word matching.

The natural-kind optimist calls upon us to collaborate with lexicographers, linguists, psychologists, and anyone else who might be able to marshal data. The gruesome pessimist advocates a benign quietism with respect to any departures from clear cases. There need not be a correct answer forthcoming, or else we are utterly unable to know what it is.

---

[70] Note that the gruesome pessimist can still agree that the lexeme conception is in some interesting sense the theoretically deepest, since the latter point does not entail anything like the claim that criteria of identity will be forthcoming.

Now, the appearance of the category of lexemes in theoretical linguistics may seem to offer hope to the natural-kind optimist. But such hope may well be premature. When one looks at discussions of lexemes within theoretical inquiry, confident judgments of identity and difference—and associated diagnostics—are typically local. For example, the tests we have looked to for a single lexeme in the case of polysemy are designed to apply to a temporally and communicatively local set of uses. When it comes to performances that are temporally and/or spatially distal we get little systematic guidance. Perhaps there is a case to be made for natural-kind optimism, but most of the work remains to be done. The actual scenario may well be bleaker: there are no joints in nature that can resolve distal questions of word identity one way or another.

There are two subspecies of the gruesome pessimist. One is the sloppy realist. On this picture, the unsettled questions turn out to rest on borderline cases and are to be handled using the correct theory of vagueness (whether it be epistemicist, supervaluationist, or whatever). In that case, there either are facts we may never know or simply no facts at all about the myriad borderline cases left unresolved by our capacity to settle questions in the area.

Assuming that rampant semantic indecision does not mandate repudiation of this or that ontological posit, we are left with a practice that, while not so bad as to warrant skepticism, is too flimsy to warrant sustained metaphysical investigation. On this perspective, there is good news and bad news. The good news is that the elusiveness of questions of word individuation need not indict our practice of positing words. The bad news is that the accessible facts about words run so shallow that there is little philosophical payoff to ruminations about word identity. Those who pursue questions of word individuation and hope for systematic answers are almost invariably in the grip of a faulty picture of the semantic mechanisms that underlie thought and talk about words.

A second version—atheism—*denies* the existence of words. Words are a mistaken ontological projection. There are interesting relations of connectedness among performances, but no interesting equivalence relation. A failure to recognize this has led to an unwarranted postulation of objects whose coherence is ultimately undone by the lack of grounding equivalence relations. (It is as if we posited directions in a setting where it turned out that *being parallel to* is not an equivalence relation.) Note that the view of words as continuants does not fare any better here. If there are no interesting equivalence relations of connectedness, then there is no useful way of gathering performances into continuants either.[71]

The defense strategy for the sloppy realist is clear enough. She will emphasize that words are no worse off than songs, dances, and so on. Unless some general atheism about the posits of common sense can be motivated, then atheism in this case is likely to be an overreaction. Is there a productive way forward for the atheist? One promising way may be to focus on "fission" cases. Suppose x belongs to a community that uses a particular word, 'happy'. Two communities c_1 and c_2 pass by x's community and, by x's lights, appear to pick up that word and return to their homelands with it. x has a description of this case that by her lights is extremely natural. c_1 picks up that word and comes to pronounce it in one way, while c_2 picks up the word but comes to pronounce it in a very different way—let us say 'harpy' and 'hapry'. Suppose c_1 and c_2 come to attach different meanings to the relevant words. Again, this will have a natural description by the lights of x: c_1 and c_2 use the same word with slightly different meanings. Suppose c_1 and c_2 encounter each other. The c_1 members treat the descendents of 'happy' in c_2 as articulations of a different word than the descendents of 'happy' in c_1. Similarly for c_2. But a natural perspective on the situation by the lights of x will be that c_1 users and c_2 users do not realize they are using the same word. Indeed, very natural considerations will *force* such a conclusion. After all, it is eminently natural to think that c_1 is using the same word as the original, albeit with a different meaning, and that the same is true of c_2. Given these facts, x will be forced to think that the two communities are using the same word without realizing it. But now think of things from the perspectives of c_1 and c_2. Suppose the c_1 users pick up the c_2's uses of 'hapry' and the c_2 users pick up the c_1 uses of 'harpy'. Members of both c_1 and c_2 will find it natural to think they are using two different words. Even if they learn that there is a common origin, this likely will not affect that judgment. The situation will be similar to that of the English 'moan' and 'mean', which seem obviously to be two words despite a common origin in Anglo-Saxon ancestry.

So who is right? x finds the claim 'they are both using the same word' to be obviously true. The c_1 and c_2 users find the claim 'we are both using different words' to be obviously true. Reckoning both judgments to be borderline is very disturbing, for it seems to imply that our own judgment that 'moan' and 'mean' are different words is at best borderline. But siding with one perspective seems bizarrely chauvinistic. Atheism offers a way out of the dilemma: there can be a satisfying reconciliation if only they abandon the ideology of word identity for that of connectedness. All parties can agree (on at least one natural construal of connectedness) that the c_1 uses are connected to the x uses, that the c_2 uses are connected to the x uses, but that the c_1 uses are not connected to the c_2 uses. But what reconciliation does the sloppy realist have to offer? She might look for guidance at what metaphysicians have to say about other fission cases. Where there is person fission, some metaphysicians say there were two people all along. But the judgment that there are two words all along in the case just described seems almost as disturbing to the natural conception as atheism.[72] Given the prima facie conflict with common sense, atheism admittedly faces an uphill battle. But it is far from clear that it is a battle that cannot be won.

---

[71] This strategy adapts what Kit Fine has claimed about meanings to the level of words themselves. See Fine, *Semantic Relationism* (Malden, MA: Blackwell, 2007).

In conclusion, let us turn briefly to the question with which we began this section. Given the polysemous nature of the word 'word', there is no context-independent answer to the question "What are words?" any more than there is a context-independent answer to the question "What are books?" Nevertheless, Kaplan was onto something. There are many contexts in which the word 'word' picks out neither particular performances nor an abstract entity which merely serves to encode the superficial form of particular performances.[73] That said, it is doubtful that the best treatment of such uses adopts a conception of words as four-dimensional continuants. Rather, there is more hope for a model of words as abstracta, though one that breaks with the standard type-token model's picture of the relevant abstracta as pattern-like. The challenge remains to provide a criterion for word matching between performances. Having eschewed superficial criteria, this challenge takes on a forbidding character, since there is no obvious surrogate that can provide a criterion once the form-theoretic ones have been dispensed with. Rather than expect that such criteria will be forthcoming, we must take seriously a conception of our practice which guardedly endorses an ontology of words while despairing of such criteria. Whether the reflective metaphysician ought to sign on to such an ontology as well is a vexed question. But whether one opts for sloppy realism or atheism, one thing is clear. As far as individuation criteria are concerned, a lexeme-like conception of words, even when supplemented with the tools of theoretical linguistics, may have just the wrong mixture of superficiality and indecision.

---

[72] One possible way out for sloppy realism is a sneaky contextualism. The c_1 and x users pick out a different domain by 'word'. Such maneuvers seem, prima facie, rather desperate.

[73] Indeed, the context of this sentence—where we use the expression "the word 'word'"—is one such context.

JOHN HAWTHORNE

University of Oxford

ERNEST LEPORE

Rutgers University

COMMENTS AND CRITICISM

WHAT ARE WORDS? COMMENTS ON KAPLAN (1990), ON HAWTHORNE AND LEPORE, AND ON THE ISSUE[*]

David Kaplan in his 1990 paper "Words," and John Hawthorne and Ernest Lepore in theirs in this issue, address three tightly intertwined questions about words in general, and about names in particular: "What are they?" "How can (or should) they be individuated?" "When are two instances instances of the same one?" In these comments I will concentrate on the first question and only as it pertains to words in general, not names in particular; that is, I will focus on the question "What are words?" and leave for another occasion the consequences of what I write for the other issues they raise.

The question "What are words?" is woefully underspecified. Its only clues are the topic 'word', the default interrogative 'what', and the default plural tense bearer. Reformulating it, as Hawthorne and Lepore do at one point, as "What is the nature of words?" hardly adds anything helpful. Do words even have a nature? The question can be understood as querying for a genus under which words can be enlighteningly subsumed, and then for how they differ from other members of that genus, and from things not under the genus at all. But what could such a genus be? Sounds? That leaves out inner speech, writing, and signs. How about illusions? After all, utterances (inner and overt), inscriptions, and signs are largely illusory: their linguistic attributes are not in the physical signal but are assigned in top-down perceptual processing. (Consider, for example, word boundaries in flowing speech, stress in writing, filled reduced vowels, meanings, functional relationships, and so on. Think of how foreign speech comes across.) So utterances—and for that matter inscriptions and Sign languages productions—group naturally with the Big Dipper, the Müller-Lyer line, the enlarged moon near the horizon, and the

---

[*] An earlier version of these comments was presented at the 2010 meeting of the Pacific Division of the American Philosophical Association. At that meeting David Kaplan read a response to Hawthorne and Lepore's paper, but I was never given a written copy of that response and was not able to absorb much of it from Kaplan's animated presentation while I was setting myself to give my own talk. Nor did I have a chance to review the version included in this issue, which was sent to the JOURNAL long after my own paper had been set in proof form. So, in these comments, I limit myself, as far as Kaplan in concerned, to his 1990 paper. Hawthorne and Lepore had sent me a copy of their paper well in advance of the meetings, and I am most grateful for that.

0022-362X/11/0809/486–503

missing fundamentals in violin performances, as perceptual illusions. We could—maybe should—treat the question as calling for how perceived words are like and unlike other illusions. That would be an intriguing inquiry indeed, but it is not the one that engages our authors. They do not think of individual utterances or inscriptions as illusions at all. They think of them as denumerable physical entities, and they think of words as distinct further denumerable entities with spatiotemporal instances.

But that hardly pins down the question "What are words?"; nor does it pin down what is at stake in their debate. And so, not surprisingly, Kaplan on one side, and Hawthorne and Lepore on the other, approach the question and its requirements with very different concerns in mind and end up often at cross purposes, leaving some essential aspects of words unattended.

I. KAPLAN'S TREATMENT OF "WHAT ARE WORDS?" IN "WORDS"

Kaplan describes his undertaking in his paper as follows: "I propose a...model according to which utterances and inscriptions are *stages of* words, which are the *continuants* made up of these interpersonal stages along with some more mysterious *intra*personal stages" (Kaplan's italics).[1] He does not propose an analysis of whatever concept, if any, the English word 'word' might encode. Nor is he after the satisfaction conditions for the predication '...is a word'. Nor is he after a genus-differentia determination of distinctive properties that only words might have in common. Nor is he after smoking out whatever ontological presuppositions might underlie common talk about words. He is after a "model," a "different conception of the fundamental elements of language."[2] He is thus proposing to displace our everyday, unanalyzed, multiple notions of words, as well as the orthodox Peircian type-token doctrine, and possibly other notions, and to replace them with a presumably more useful new notion embedded in a new theoretical framework.

His proposal is that we think of each word as what some philosophers used to call a substance. Words are to be thought of as persistent individuated entities each of which retains its identity as it grows and develops through what Kaplan calls "stages." The term 'stage' is used idiosyncratically. What he calls stages are actual physical utterances and inscriptions produced at specific times and places by specific individuals. They are not what etymologists might mean by the "stages" of a word, namely, successive form and meaning combinations arising in a process of historical evolution. Each word in Kaplan's "model" is a growing tree-like structure made up of actual utterances and inscriptions, each of which is related to a predecessor in the structure, a predecessor that the speaker or writer intends to copy, though not necessarily exactly. Such predecessors may themselves be actual utterances or inscriptions, but they may also be icons in memory. Furthermore, each word, in Kaplan's "model," is rooted in an initial stage, its first production in time. That stage does not stand as intended copy to any predecessor. Its role is to individuate the word containing its successive repetitions. Hawthorne and Lepore construe Kaplan as also viewing these structures as mereological joints of their stages, but I am not sure that he is actually committed to such a view or its shortcomings.

---

[1] David Kaplan, "Words," *Aristotelian Society Supplementary Volume*, LXIV (1990): 93–119, at p. 98.

[2] *Ibid.*

Kaplan's invitation to change how we think of words calls not for empirical evidence, obviously, but for recommendations. What is there to recommend it? A number of things are mentioned explicitly in the paper.

His "model" does not have the defects of the classical type-token conception. That conception—according to Kaplan—entails that each word (type) is an abstract template matched exactly by each of its instances (tokens). It is useless, since many, if not most, utterances of a word are dissimilar in some respect and thus could not all mirror exactly one common template.

His "model" furthermore does justice to the fact that words, whatever they ultimately turn out to be, are at bottom human creations, and that they are mostly acquired through interpersonal exposure.

His "model" also allows us to hang on to many platitudes of ordinary talk about words. For instance, that each word has an origin and has undergone and will in all likelihood continue to undergo changes of meaning, pronunciation, and spelling, in short has a history; that many words are pronounced and spelled differently by different users because of dialectical differences, circumstances, or other contingencies.

But most important, according to Kaplan, his "model" includes a principle of individuation for words in general, and for names in particular, that allows for numerically distinct words that are nevertheless qualitatively identical (within tolerable limits) in pronunciation, spelling, and direct reference qua meaning. That is not a platitude! It is a consequence to be prized because, according to Kaplan, it allegedly yields a solution to the Paderewski conundrum. Under Kaplan's "model," misinformed Peter can allegedly believe that Paderewski is a musician and also believe that Paderewski is not a musician without being guilty of (knowingly) subscribing to a contradiction. Furthermore, this can be said of him without imputing to him contradictory beliefs (not the same thing: the latter pertains to the effability of the imputation; the former, to the belief state itself).

I must admit that, in spite of sincere efforts, I was unable to reconstruct Kaplan's solution. The paper describes two sets of circumstances under which two words (or at least two names) can be numerically distinct though qualitatively, that is, semantically, phonologically, orthographically, indistinguishable. The first description is explicitly put forth as solving the Paderewski conundrum; the other is part of the discussion of proper names in their distinct guise as generic names and as common currency names. However, neither extricates us in any obvious way from the conundrum. The first requires that a memory icon in Peter's mind/brain accidentally split at some point in his life. Even if we disregard that there is no evidence that such splitting of memory icons, or traces, or representations, is even possible, the Paderewski conundrum rests on an epistemic predicament that does not hinge on such a possibility. It suffices that Peter heard two unrelated reports about Paderewski that he never connected. The second description requires that some individual be baptized on two different occasions (possibly by two different people) with the same generic name. The occurrence of such an episode is more plausible than the occurrence of splitting memory icons, but the Paderewski conundrum does not hang on such baptismal episodes.

Hawthorne and Lepore read Kaplan as being more committed to tenets of folk linguistics than I do, but some of their more telling objections do not depend on this. Kaplan's proposal, after all, includes some empirically testable ideas. Their most decisive objection[3] is that Kaplan's "model" and principle of individuation require that each time we use a term, we use it with the intention of mimicking a specific previous use or a memory icon of that term, or that we create a new word, and thus that we never access the generative powers of our morphology. And if that is false—which it indubitably is—then there is little left of most of Kaplan's so-called stages, and thus little left to make up words as he thinks of them. It would, in fact, be surprising if there were more than a small number of such words. Maybe a few names. And even these will turn out to behave in ways not covered by his "model."[4]

My own strongest objection to the model is of a different order. Kaplan's "model" is too barren to be adopted. Any reasonably insightful and useful conception of words would do more than accommodate commonplaces about words or the fact that people occasionally mimic perceptually and referentially similar performances, or speculate imaginatively about mental contingencies, or raise hopes to solve a vexing conundrum. Any worthwhile conception would take aboard that words function as constituents of phrases and sentences. It would acknowledge that they play their defining roles merged with other terms, and thus, that—whatever their intrinsic perceptual and referential features—it is of the essence of words that they can appear in juxtapositions through which they receive and assign thematic roles,[5] and stand in various functional relationships. It would make room for such facts as that the first word in 'John is easy to please' denotes a theme, whereas in 'John is eager to please' it denotes an agent; that a different difference holds between the occurrences of 'tomatoes' in 'I want to grow tomatoes' and in 'I want tomatoes to grow'; that 'dog' and 'dogs' are related in rule-governed ways complicated by pairs like 'mouse' and 'mice'. Any worthwhile conception of words would acknowledge that our ability to parse continuous speech into words itself calls for explanation, as does the fact that young children can come to do this in their linguistic developments. And so on, and so on. Kaplan's "model" meets none of these basic essential desiderata.[6]

---

[3] Some of their other objections are discussed in the next section

[4] See Ora Matushansky, "On the Linguistic Complexity of Proper Names," *Linguistics and Philosophy*, xxxi, 5 (October 2008): 573–627.

In short, Kaplan's proposal, as it now stands, rests on too truncated and impoverished a view of lexical mastery. But my objection is not that it failed to provide a fuller account of that mastery. That was not Kaplan's project, and that would be an unreasonable demand. My objection is that his proposal distorts and diminishes that mastery, and is therefore incompatible with any plausible account of that mastery. His proposal is original, is ingenious, is intellectually adventurous, is addressed to a real set of issues, probably describes real spatiotemporal entities (unnamed before him, and of no linguistic interest, but no doubt acceptable to any tolerant ontology); it may even hint at a way out of the Paderewski conundrum, but it is just not a trustworthy starter for thinking about words or their identity.

II. HAWTHORNE AND LEPORE'S TREATMENT OF "WHAT ARE WORDS?" IN "ON WORDS"

Unlike Kaplan, Hawthorne and Lepore are not motivated primarily by a specific philosophic enigma. They are concerned primarily with certain uses of the English word 'word'. Though not explicitly, they interpret the question as calling for a Quinean "semantic ascent" on that word and those uses. Mindful that the English word 'word' is polysemous, they aim at uses they deem to be at the heart of Kaplan's concerns, uses they assume—though without offering any evidence—philosophers and linguists rely on in connection with "a variety of puzzles and problems that are foundational to the very subject matter of philosophy of language and linguistics,"[7] uses in "many contexts in which the word 'word' picks out neither particular performances nor an abstract entity which merely serves to encode the superficial form of particular performances."[8] They assume—as I read them—that the English word 'word'—when so used—is used to refer and has an extension. The question for them becomes "What is peculiar to the members of that extension?" "How are these members individuated?" They favor a familiar answer to the former: "...the *abstracta-articulations model*."[9] "[T]here is more hope for a model of words as abstracta, though one that breaks with the standard type-token model's picture of the relevant abstracta as pattern-like."[10] And they hold that no unproblematic answer is to be had to the second question.

---

[5] This is not to hold that meaning is primarily contextual.

[6] More about this in the last section of these comments.

Whatever the merits or defects of their "model"—which they present as more valid than Kaplan's—their paper is not primarily devoted to its defense, but is primarily designed as an evaluation of Kaplan's "model." I will confine my discussion to that evaluation. I take it to rest crucially on their presumption—mistaken, I think—that Kaplan intended to explicate certain uses of the English word 'word' rather than advocate an altogether new concept. I will not dwell on that point, but will adopt that interpretation of "What are words?" for the sake of discussion. I will dwell, however, on their further presumption that certain manners of speaking about words are authoritative about what words are. That presumption, at least as it is treated in this paper, is blind to too many essential features of words. I will concentrate on three early typical passages of the paper that by themselves are sufficient to bring out why the proposals in "On Words" will not do.

The first of these passages is the passage in which Hawthorne and Lepore discuss a thesis about which they agree with Kaplan. As they put it "...we quite agree with Kaplan that a philosophically satisfying theory of words cannot proceed entirely within a shape- or form-theoretic framework."[11] They hedge the point in two ways, first by inserting 'entirely', and second by remarking shortly thereafter: "We do not deny that there may [*sic*] be *certain* contexts in which we use 'word' to speak about entities that are individuated by shape or form" (their italics).[12] They then write that they are not interested in those uses. But even with those somewhat puzzling hedges, the thesis—and the reasons they give to bolster it—call for close critical scrutiny. For what is at stake are serious theoretical options that the claim would close, but that a sound philosophic approach should keep open and explore.

---

[7] Hawthorne and Lepore, "On Words," this JOURNAL, CVIII, 9 (September 2011): 447–85, at pp. 447–48.

[8] *Ibid.*, p. 484.

[9] *Ibid.*, p. 453.

[10] *Ibid.*, p. 484.

[11] *Ibid.*, p. 451.

The thesis is initially brought up by Kaplan as he argues—as I mentioned above—that words could not be a kind of (abstract) template duplicated by each and every one of their instances. For that thesis would entail that all the physical instances of any given word would also duplicate each other, and it is obvious that they mostly do not, and need not for practical purposes. That much is beyond dispute. In fact, we can often predict how they will not. Take, for instance, the ordinary preposition 'in'. In normal speech, the last segment (the segment spelled 'n') is produced and pronounced one way when followed by, for instance, 'New York' (voiced, noncontinuant, alveolar, coronal, nasal, sonorant, ... n), another way when followed by, for example, 'Paris' (..., noncoronal, bilabial m), and yet a third way when followed by, for example, 'Kansas' (..., noncoronal, velar eng ŋ). Note that a relatively explicit single notion of form is at hand in such instances: acoustic form.[13] Analogous observations can be made about written tokens, where the relative notion of range of forms, and their interdependencies, is also readily analyzed.

None of this establishes, however, that a good, insightful, ontologically sound theory of words cannot "proceed...within a shape- or form-theoretic framework." It shows at best that such a framework requires refinements. And, of course, no framework pertaining to forms alone can hold the full story about what is essential to words: words have other essential characteristics. But that is not part of Hawthorne and Lepore's argument.

Hawthorne and Lepore do not limit their evidence, as does Kaplan, to the fact that spoken tokens of a word can differ from each other, or that written tokens of a word can differ from each other. They draw on more drastic differences. They write, "Simple reflection on the fact that the same word can be written, uttered, signed, Brailled, or semaphored already renders such a conception dubious: after all, there is hardly anything shape- or form-like in common among an utterance, an inscription, a hand gesture, and a bump on a panel."[14] They might have pointed out further that there is hardly anything shape- or form-like in common between the oral gestures performed in the production of a word and the acoustic events that result, let alone the auditory experiences induced, or between the latter and inner spoken words, or words in memory, or between any of those and the kinds of instances they mention.

---

[12] *Ibid.*

[13] I say acoustic form, not auditory form, because we are apt not to perceive these differences except under special conditions We are more attuned to invariance in formant patterns than actual frequencies and amplitudes. Of course, articulatory pattern is also implicated, but that is not relevant for their discussion.

No doubt simple reflection on these facts may cast doubt on what they call the form conception, but less simple reflection should not, but elicit that these facts carry different implications when scrutinized more closely and put together with further facts.

Consider first their assertion that "the same word can be...signed...." That may come across as an innocuous remark, but it is not; it is symptomatic of an outlook on language too removed from reality to warrant any philosophic claim, let alone one about methodology. In a footnote they specify that by "signed" they mean signed in something called "Signed English"; they do not mean American Sign Language (ASL). And there are indeed signing systems called Signed English, also called Manually Coded English (MCE)—a number of them. But none of these systems, in spite of their name, is an encoding of English: each encodes a form of pidgin ASL.[15] The basic vocabulary of these systems is largely identical to the vocabulary of ASL, with some additions for inflections, articles, pronouns, copulas, affixes signed in ASL manual alphabet abbreviations, and simplified uses of space and indexing. The very morphology is thus utterly unlike that of English. For instance, the sign to express "believe" is a compound of "think" and "marry." And, of course, the analogue of phonology relies on completely unrelated structures of features; the syllables do not begin to map onto English syllables; and so on. The surface word order and the distribution of lexical items does come closer to that of English than that of ASL. And not by accident. The systems were created to help communication between hearers and Deafs, and to help Deafs parse written English.[16] However, syntactic word surface order and distribution presuppose, but do not exhaust the ontological substance of words, which is the topic under discussion.[17] MCEs do not

[14] *Ibid.*

[15] One is in fact called "Pidgin ASL," a label assigned, I believe, by Signers.

[16] Actually, many sentences in Signed English are indistinguishable from their ASL version.

[17] English can no more be signed than it can be Frenched. "Je fais pas aimer brocoli" is not Frenched English for "I don't like broccoli." There is no such thing, though we might find a use for the phrase "Frenched English" as we have for "Signed English."

even encode a natural language (a language that can be acquired as first language), let alone English. The possibility of relying on English syntax while using the ASL lexicon raises interesting issues that do deserve philosophic attention, but that is largely because no same word can be both uttered and signed.[18]

Let me turn next to the remainder of Hawthorne and Lepore's presumed evidence, namely, that there is hardly anything shape- or form-like in common among the various other types of instances on their list. That indisputable fact does not entail that a good, insightful, ontologically sound "theory of words cannot proceed...within a shape- or form-theoretic framework." It only entails that such a theory could not itself be simple, and that it would have to be specific about distinctions and take care of details. It would take into account explicitly that spoken and written instances belong to distinct ontological domains, domains each grounded on a specific range of essential attributes. It would indeed take into account that spoken tokens have articulatory, auditory, and acoustic properties; that their elements vanish immediately upon production; that they occur sequenced in time; that they are produced by mouth; and so on. It would also take into account that none of this holds for written tokens, and that written tokens, unlike spoken ones, have topological, material, and optical properties, as well as that the elements of written tokens persist in time, are linearized in space, and are produced through many different processes.[19] It would also crucially take into account that written words are cognitively parasitical on spoken ones, which they presuppose and encode (semaphored and Morsed words, in turn, are parasitical on written ones), and thus that though spoken ones can be acquired through exposure, written ones require prior ability to speak.[20] Such a theory would take into account that even within the realm of spoken words, different types of forms interact: forms belonging to sequences of articulatory gestures, forms belonging to auditory experiences, forms belonging to inner speech, underlying forms, and so on, and so on. But, most importantly, such a theory would also allow for the fact that these profound differences between forms are bridged by principle-governed, analyzable interdependences and mappings (some conventional). So, for instance, chronological

---

[18] There is such a thing as the Rochester method, which relies heavily on alphabet signing, but it is basically unusable, and my comments below about writing and semaphore apply to the uses of that method.

[19] Inner speech, which I leave out, has its own peculiar ontological dimensions.

[20] This has to be qualified in some respects, but not in ways that matter for present purposes. See for instance Stanislas Dehaene, *Reading in the Brain: The Science and Evolution of a Human Invention* (New York: Penguin, 2009), chapter 2.

sequencing in speech is mapped in linear spatial order in writing; phonemic segmentation is mapped in linear segmentation; phonological attributes are mapped on orthographic ones; and so on. Furthermore, it would establish whether—as seems most plausible—for each set of interrelated instances there exists a core instance to which all of the others, in spite of their ontological dissimilarities, are related in systematic ways. The dimensions and peculiarities of such core instances raise knotty questions that beg for cooperation by philosophers, linguists, cognitive scientists, and neurologists, but towards whose solution progress is actually being made. (I will come back to this point in the last section.) In any case, conjectures about such core instances and their form, or theorizing about the reality of inter-modality connections, must not be casually dismissed because their existence is not presupposed by certain manners of speaking. Manners of speaking are no doubt convenient for many purposes, but they are not canonical about facts or even plausibility. Nor should "shape- or form-theoretic frameworks" be dismissed on the false presumption that resemblance—either exact or vague—is the only relation through which the forms of word manifestations can be systematically mapped on each other. Regardless of how we might allude to the facts, simple reflection on the fluent, spontaneous performances of speaker-hearer-writer-readers should render such a presumption extremely dubious.

I next turn to another passage a short way later in "On Words." Unlike the one discussed above, it states an objection to Kaplan's proposal. The objection is that "Kaplan is committed to the view that all words are articulated in one way or another."[21] In other words, that Kaplan's model has no provisions for words that have no instances, have no performances, have not even an initial occurrence. They put forth essentially three considerations in support of the view that there are such words. First, it is only a contingent fact that, for instance, some affixations of the English prefix 'un-' and some arbitrary English adjective has never been produced. They might have been produced. Second, let *n* be the largest number of occurrences of 'anti-' in any instance of 'anti-...ballistic' ever produced. An instance with one more prefixed 'anti-' could have been produced, even though no such instance happens to have ever been produced. That too is but an accident. The relevant morphology is productive and open-ended. Third, more generally, the productive rules of morphology governing affixation and compounding are mute about which affixations and compounds have actually been produced.

---

[21] Hawthorne and Lepore, *op. cit.*, pp. 454–55.

And so, to use their formulation, they urge, it would be very awkward to deny that "such words" exist.

Each of these observations is beyond dispute. And the awkwardness of any unqualified blunt denial that such words (never-produced words) exist would probably be real. We do not converse that sparsely, as Grice has taught us. But to infer that there are therefore abstract words with no instances, that the furniture of reality, so to say, includes such abstract entities, is a plain non sequitur. What there is are human mind-brains endowed with the capacity to produce utterances (and inscriptions and so on), and that capacity is not exhausted by past, present, or future productions. The above observations do not entail that there are such things as words with no instances. They do not presupose an ontology of instance-less abstract words. Without speaker-writer mind-brains there would be no instances at all. With speaker-writer mind-brains there still are no noninstantiated words, though there are unexhausted capacities. In short, speaker-writer mind-brains endowed with grammars and lexicons leave no need for abstracta. Nor should any philosophic theory posit them out of respect for how we sometimes economically express the facts.[22]

The last passage I shall discuss before turning to more general considerations is about what they call the "Evolutionary Constraint," a proposition on which they again agree with Kaplan and which they advocate as a "framing principle," that is, a principle that the metaphysics of words should respect, presumably because they hold that it is metaphysically sound and important. The principle is, "As it is transmitted from one epoch to another, [a word] can change not only its phonological and orthographic contours, but its semantic and grammatical ones as well."[23]

How sound is the principle as stated? The facts about the semantics and syntactic "contour" of a word touch on too many controversial technical issues to be fruitfully examined here. But what I write below about the phonological changes should apply with some nonessential modifications to them. The phonological facts, and the phonetic ones they induce, are more readily describable.

Normally, single words do not change in isolation, but whole families of words that share features change together as certain shared constituent features get replaced in shared phonological environments. These changes are often remarkably law-like[24] and are often labeled as such by linguists. So we have Grimm's law, responsible for the spiritization of voiceless stops in Germanic; Verner's law, which followed it in time and effected voicing in certain configurations; the Great Vowel Shift in English in the fifteenth century, resulting in drastic but systematic changes in long vowels; the current North American Vowel Shift; and the Valley Girls Rise, which is spreading from California through the United States and is driving my wife crazy. It was the reconstruction of such systematic featural changes that led to the revolutionary discovery of Indo-European and pre-German, otherwise completely unattested languages.

---

[22] For a splendid set of arguments for the contrary view, see Linda Wetzel, *Types and Tokens. On Abstract Objects* (Cambridge: MIT, 2009).

[23] Hawthorne and Lepore, *op. cit.*, p. 459.

[24] They are also remarkable similar in form to synchronic regularities.

Furthermore, the only changes that necessarily and sufficiently made up those events were people changes. People changed, not "words"! Nothing like the disembodied Platonic words alluded to in Hawthorne and Lepore's "Evolutionary Constraint" was involved.[25] And the same is true for all diachronic lexical changes past, present, and future.[26] What actually happens is that specific, real, flesh-and-blood, human individuals, often across generations, often influenced by contacts with foreign neighbors or conquerors, or moved by pressures for articulatory and auditory efficiency, or the wish to follow some fashion, or other motives or pressures, start to talk differently, to move their articulators in different patterns, to internalize different rules or constraints, sometimes to the point of achieving creolization, and to adjust their perceptual expectations accordingly. In the case of Grimm's law, for instance, Germanic speakers of one generation produced bilabial stops in certain contexts, and then Germanic speakers of a later generation started to produce labio-dental continuants instead. (Think of the difference between how the Latin 'pater' and the English 'father' are produced.) Morris Halle and I have speculated about the form such changes take in the internal grammars of individuals. Whether we were right or wrong, the indisputable fact remains that talk of words changing is at best a convenient shortcut, and a way of being noncommittal about empirical details. But the moral, once again, is that anyone seriously interested in the ontology of language ought not to take *façons de parler* about change at face value. The "Evolutionary Constraint" as a guide to ontology is vacuous.

In the last section of their paper, Hawthorne and Lepore switch from issues about the defining character of words—what distinguishes them as words from other things—to closely related issues about word individuation—what distinguishes individual words from each other.[27] They base their discussion—they write—on a reading of the word 'word' "according to which it is used in a way that conforms to the lexeme conception."[28] Unfortunately they do not tell us which of the many versions of that conception and term they have in mind; how the conception they have in mind covers suppletion, or handles words in agglunative and polysynthetic languages (where the demarcation between words and phrases can be problematic); or what notions it assumes of the relation between phonology and phonetics, or between the syntax and morphology. So their switch in terminology is of little help, for the notion of lexeme raises as many thorny issues as (if not more than) our everyday notion of word. At one point they seem to admit as much.[29] After a careful, and in many ways insightful, critical review of options, they conclude that no useful criterion of identity is available now or is likely to be in the near future. And even if one is ever to be found "most of the work remains to be done."[30] They put the moral they draw as follows: "The good news is that the elusiveness of questions of word individuation need not indict our practice of positing words. The bad news is that the accessible facts about words run so shallow that there is little philosophical payoff to ruminations about word identity."[31] Their "good news" has a certain plausibility, but it leaves us in the dark as to whether "our practice of positing words" needs to come down to more than just using the word 'word' as a syntactic noun. Their "bad news," on the other hand, advocates a resignation that should be resisted. The ontological issues are real; they refer to important gaps in our image of reality, of lexical memory, of language capacities in general, and of ourselves. Though these are thorny, labyrinthine issues, and though their solution may call for widespread explorations, we owe it to ourselves and to those who rely on us to pursue them. Fortunately, the "accessible facts about words" do not run shallow at all, except for those who limit themselves to commonplaces. In what follows I turn to that aspect of things.

---

[25] If they are abstract entities, how could they change?

[26] What are we then to make of etymology? Etymology is not ontology. Though it abstracts from facts, it is not about abstracta. Actually, the methodology of etymology deserves more attention from philosophers than it has received.

III. FURTHER THOUGHTS ON THE QUESTION "WHAT ARE WORDS?"

As I wrote at the outset, "What are words?" is woefully underspecified. Kaplan reads it one way; Hawthorne and Lepore read it in a radically different way. Neither reading turned out to be very fruitful. In this section I want to sketch a reading that strikes me as more promising and more urgent for the ontology of words and linguistic constituents in general.

---

[27] They actually discuss jointly the two questions "How can words be individuated?" and "When are two instances instances of the same word?"

[28] Hawthorne and Lepore, *op. cit.*, p. 476.

[29] They write, "When one looks at discussions of lexemes within theoretical inquiry, confident judgments of identity and difference—and associated diagnostics—are typically local." *Ibid.*, p. 482.

[30] *Ibid.*

[31] *Ibid.*

Contrary to what Hawthorne and Lepore write as quoted above, the accessible facts about words do not run shallow. On the contrary, we know a lot about words. That is not surprising: they have been under intense study since at least the fourth century BC (Pāṇini). Here is a random sample of things we know about words. Every human natural language, no matter how unrelated, systematically organizes words.[32] Each word, in every language, is constituted of hierarchies of segments and belongs to lexicons made up of other words, or is derived or compounded from such other words. Words are distinguished by their syntactic, phonological, and semantic features,[33] which affect their distribution in phrases, for they get merged with other words into phrases and sentences under universal (though sometimes idiosyncratic) and parameterized rules or constraints. Words are the domain of morphological conditions on affixation, clitization, and compounding. They are the domain (depending on the language) of stress or tone assignment.[34] They are the theme of phonological and morphological processes of assimilation, vowel shift, metathesis, reduplication, and phonotactic constraints. They manifest themselves as sequences of articulatory gestures and auditory events or sub-vocal performances by specific individuals at specific times. For some people, such episodes can be guided or encoded by alphabetic, syllabary, or ideographic scripts. New words can be acquired long after all other aspects of one's language have been fixed. And so on, and so on.

These facts—and facts like them—clearly raise ontological issues expressible as "What are the truth makers of these facts and facts like them?" Or, somewhat more controversially, "What sorts of things must exist in a world in which they obtain, but need not exist in a world in which they do not obtain, and why?" And these are also reasonably promising specifications of our question "What are words?"

One thing is certain: these facts are rooted in people. They hold because certain facts hold of people, flesh-and-blood speaker-hearers, not of abstracta in some Platonic heaven or grinny Wonderland.

---

[32] Whether the notion is usefully extended to cover the smallest free forms in linearized artificial notational systems (for example, logics) that have no phonology is debatable, and is probably best set aside until what is at stake about words becomes clearer.

[33] That needs some pedantic qualification; for there are pleonasms, empty categories, and so on.

[34] That too needs some pedantic qualifications.

Furthermore, they hold because a person who has internalized words is thereby in a state that enables that person to meaningfully produce, parse, understand, and entertain utterances—a very special, probably species-restricted capacity. But it is also a capacity that requires more than the acquisition of individual words. It requires the possession of a whole grammar, that is, a phonology, a morphology, a syntax, a semantics, a pragmatics, and so on, not to mention a certain anatomy and specialized neuro-physiology. It is, to say the least, a very complex capacity! There can be no words where there is no such capacity.

That capacity includes the capacity to unconsciously plan and execute complex processes, as attested by the fact that even in informal, casual talk or thought, our sentences flow out preplanned, structured, conforming to all manner of functional and linearization requirements and constraints. Subjects come before verbs in English; 'wh-'s occur in complementizer positions and not where they get their thematic role; 'do' in English picks up tenses where main verbs are blocked from doing so by negations (or interrogation); agreements are respected; affixes assimilate phonetically to their stems (or vice versa); and so on, and so on.[35]

As we speak (and listen) even casually, we thus carry out complicated processes. Not merely cognitive processes, but physical ones as well that implicate motor neurons controlling our abdominal muscles, vocal folds, velum, back of the tongue, blade of the tongue, sides of the tongue, and lips in incredibly subtle, targeted gymnastics.[36]

Yet none of these processes, not even their input (what initiates what we are going to say), is accessible to awareness. None of that planning, none of these physico-mental, presumably intentional activities are open to our consciousness! That is an aspect of hearer-speaker predicaments that deserves much more attention than it has received from philosophers. We are held responsible for what we say. We hold ourselves responsible for and in charge of what we say and how we say it. We impute all manners of reasoning as determining what we say. And that is as it should be. We are not automata! But at what point in our talk do we start or cease to be agents?

We do not even have, as part of our everyday conversational notions, notions that are up to the task of stating, or of bringing to consciousness, what we presumably do intentionally when we speak. Let me illustrate this with a revealing conversation I had a few years ago with my lovely, then two-year-old granddaughter Eliza. It also shows how seriously flawed Kaplan's notion of repetition is.

---

[35] I am limiting my illustrations to English speech. They are readily expanded and adjusted to cover other and prima facie very different languages.

[36] Related processes are implicated in writing, reading, signing.

Eliza: Me play.
Sylvain: Eliza, sweetheart, say "I play."
Eliza: Me play.
Sylvain: No, no, not "me," "I." OK? Say "I."
Eliza: Me.
Sylvain: Say "ayayayay."
Eliza: Ayayayay.
Sylvain: Say "bye bye."
Eliza: Bye bye.
Sylvain: Great! Now say "I."
Eliza: Me.

Puzzled by this exchange, I turned to my friend and colleague Ken Wexler, who, unlike me, really knows linguistics and linguistic development. It turns out that the phenomenon is widespread and has been studied. The explanation runs roughly as follows. Up to a certain age, children do not "have case" in their inner grammar. As a consequence they do not produce nominative case. A noun phrase gets nominative case when used in a certain relationship to a tensed verb, roughly, when it is used as the subject of a tensed verb. (Genitive and accusative or dative require different syntactic relationships.) Case is phonologically and phonetically overt in English, unlike in many other languages, only on certain pronouns, though it is assigned to other noun phrases covertly as well.[57] 'I' is the nominative case form of the first-person singular pronoun in English; 'me' is what is called the "unmarked" case, the default case, the no-case, so to say. (So when someone asks, "Who did this?" you answer, "Me," not "I," in the absence of any verb.) Eliza, not having case in her grammar, produced what she did have access to: the default form. (Notice my use of the word 'form'.) And when she did acquire the full panoply of cases later in life, she was totally unaware of her transformation, and, until schooled, remained unaware of its effects. Actually case, and case assignment in agent position of tensed verbs, though an essential feature of our linguistic practice, is still not well understood and is certainly not mentioned in our commonplaces about words. But a place must be opened for it in any "philosophically satisfying theory of words."

All this puts us in a vexing intellectual predicament. "What are the truth makers of the above propositions about words?"—the way I urge that we interpret "What are words?"—must do justice to the fact that these truths hold by virtue of human-specific capacities; it must invoke the nature and character of these capacities and of the processes that realize them. That is impossible if we limit ourselves to the everyday and philosophic *façons de parler*. Hypostasizing abstracta and common currency notions only takes us in the wrong direction. Unfortunately, delving into phonology, phonetics, morphology, syntax, linguistic development, and so on is also futile at this point in our knowledge. As far as I know—and I may be simply displaying my ignorance here—no real help is forthcoming from those quarters yet. Emphasis on "yet." Emphasis on "yet." Emphasis on "yet." Hawthorne and Lepore are right when they write "most of the work remains to be done," though "most" is misleading about actual progress.

---

[57] This accounts for the fact that 'John seems to be happy' and 'It seems that John is happy' are fine, whereas 'It seems John to be happy' is not fine.

To fully spell out what I have in mind would take us far afield and require much more space and technicalities than are appropriate here. But an example from what philosophers probably deem the least problematic dimension of speech, so-called "sounds," may give an inkling of what is at stake. Phonologists study, among other things, alternations. Examples: 'electric'/'electricity', 'sane'/'sanity', 'knife'/'knives'. Members of each pair are conjectured, on good theoretical grounds, to derive from a common memorized[38] element, what is called an "underlying representation" (the word 'representation' is unfortunate[39]), through the application of rules or the implementation of constraints. So, for instance, 'electrisity' is derived from 'electrik', and in the course of the derivation the 'k' turns into an 's' sound. 'k' and 's' are specified in the theory as arrays of articulatory gestures, such as, in the case of 'k', obstructing of the vocal tract, raising of the back of the tongue, stiffening of the vocal cords. But that cannot be the content of 'k' when used as what underlies the 's' sound production in 'electrisity'. The production of 'electricity' does not involve any such gestures. But what then are we talking about when we talk about underlying phonological representations? What in the world are they? What are their features? What is their ontological and cognitive constitution? The foundations of phonology are replete with confounding issues like this one.[40]

We have been urged by Wittgenstein, Horwich, and innumerable other wise philosophers, "Do not ask for the meaning, but ask for the use." But the use of what? Underlying representations? As far as I know, the question should have us tongue-tied at this moment of our intellectual history. Emphasis: at this moment of our intellectual history!

---

[38] Not consciously.

[39] Some linguists prefer other terminology, for example, 'lexical entry'.

[40] The issue survives even if we switch from articulatory features to auditory ones.

And so, more generally, no detailed and solidly warranted answer to "What are words?"—as a question about the truth makers of linguistic facts—is forthcoming from linguistics, at least the bits of linguistics with which I have had contact. None is forthcoming at this point from the brain sciences either, as far as I know, though wonderful research is done there. But at least we are at a point where we can appreciate with some precision what we know we do not know. That is progress! Promising progress! The road to knowledge is paved with acknowledged, pinpointed ignorance. And we have good reason to believe that after more empirical discoveries, more conceptual innovations, more unification of disciplines, more theoretical developments, answers will emerge, as they have in other areas. In most likelihood, these answers will break up the notion of word into a number of more precise and theoretically manageable ones.[41] Meanwhile there is work to be done. Much is being done. Philosophers could be most helpful in that work. They have relevant analytical skills that are invaluable and that few others have had the chance to develop. But they will have to turn their metaphysical binoculars around to see the facts more closely. Not abstract facts, but concrete ones. Not how we talk about words or might prefer to talk about words to avoid puzzles, but what words come to in the actual world. As I suggested above, what is at stake is our understanding not only of words, but of ourselves.

SYLVAIN BROMBERGER

Massachusetts Institute of Technology

---

[41] For an early version of such an approach, see Anna-Maria Di Sciullo and Edwin Williams, *On the Definition of Word* (Cambridge: MIT, 1987).

COMMENTS AND CRITICISM

WORDS ON WORDS*

Hawthorne and Lepore's paper begins thus:

> In his *seminal* paper "Words," David Kaplan addresses a pair of questions that have been largely neglected by the philosophical community:
>
> (i) Under what conditions are two utterances utterances of the same word?
> (ii) What *are* words?
>
> That these questions have not received much attention is rather surprising: after all, philosophers and linguists frequently appeal to considerations about word and sentence identity in connection with a variety of puzzles and problems that are foundational to the very subject matter of philosophy of language and linguistics. Kaplan's attention to words is thus to be applauded. And there is *no doubt* that his discussion contains *many* useful *insights*.[1] [Emphasis missing in the original.]

As we say in the Passover service, *Dayenu*. That alone would have been sufficient. Unfortunately, as I have discovered to my sorrow, papers that begin like this always continue, at length, in a different vein. In fact, the very next word in their introduction is "Nevertheless."

I will respond to some of their criticisms, but I want to say at the outset that in my view, when you publish something, you put it out there for all eyes to see from their own perspective. And so, I believe

---

* This article is an attempt to reconstruct a paper delivered to a symposium at the American Philosophical Association meeting in San Francisco on April 3, 2010. It specifically aims to do no more than that. In order to avoid an endless back and forth, we all agreed to publish what was presented, rather than second thoughts based on what the others had presented. This was our deal. Hawthorne and Lepore stuck to it (noting exceptions in footnotes), and so have I, as best I can. Some undelivered comments on their Part Four were reconstructed from my notes on their original manuscript. In my own case, the mostly handwritten paper was well received, but the manuscript was lost during mutual congratulations in a bar after the symposium. I did not realize that the paper had been lost until a few months later, and by that time its contents had largely drained from my always sieve-like memory. Confirmation that it was indeed irretrievable left me in grief. Finally I understood what one of my heroes must have felt when he wrote, "Hardly anything more unwelcome can befall a scientific writer" (Frege, reacting to a difficulty raised by Russell). Perhaps I overreacted.

[1] John Hawthorne and Ernest Lepore, "On Words," this JOURNAL, CVIII, 9 (September 2011): 447–85, at pp. 447–48.

0022-362X/11/0809/504–529

© David Kaplan 2011

that all criticisms are fair. Not all are correct. But all are fair. We live in a *Rashomon* world. It even *might* be the case that *all* of their criticisms are correct. (To be more precise, there is a possible world in which all of their criticisms *are* correct. In the actual world, the actual criticisms are mostly wrong, but that does not vitiate the modal point.)

"Words" is the second in a series of three Gilbert Ryle Lectures that I gave at beautiful Trent University in Peterborough, Canada, back in 1986.[2] I was facing a problem in direct reference theory and urging that the solution depended on taking account of the differences among distinct lexical items, all directly referential and all referring to the same individual.[3] The idea of the first lecture was that because of the presumed lexical difference between names of the forms 'A' and 'B', there was already an apparent syntactical, or syntactical/lexical, difference between 'A=A' and 'A=B'. And although the semantic value of the whole needs to be calculated from the semantic value of the parts, it certainly depends on how the parts are put together, that is, the syntactical/lexical structure of the sentence in question. Truths of logic depend heavily on whether the same nonlogical constant is repeated or whether a new word appears in the second occurrence. 'Hesperus' and 'Phosphorus' are different words. In 'Hesperus is Hesperus' both places are occupied by the same lexical item, whereas in 'Hesperus is Phosphorus' the second place is occupied by a distinct lexical item. Such differences make a difference in form. I was aiming at a syntactical/lexical version of what Kit Fine has, more recently, called *coordination*.[4]

This, in a natural way, led me to consider what words are and how they are individuated, Hawthorne and Lepore's questions (i) and (ii). But there were several other motivations that led me to write "Words."

I had long been troubled by the use of what are called *quotation-names*, like

'Boston'

and

'color'.

---

[2] It was later published as David Kaplan, "Words," *Aristotelian Society Supplementary Volume*, LXIV (1990). 93–119.

[3] The first lecture, never published, was called "A Problem about Direct Reference." Kit Fine's *Semantic Relationism* (Malden, MA: Blackwell, 2007) develops a somewhat similar idea but at the semantic rather than syntactical/lexical level The third lecture, "Thoughts," was never completely written but developed the anti-Fregean thesis that sameness of meaning is more dependent on sameness of words than on sameness of a prior notion of thought. It also argued that, as followed from my "Words" lecture, Kripke's Peter had made a "linguistic error" and thus had not satisfied the requirement for disquotation.

[4] Fine, *op. cit*

What do they name? Quine had famously contrasted Boston, which he claimed contained some 800,000 people, with 'Boston', which he claimed contains only six letters.[5] (He meant six *occurrences* of letters, since there are really only five letters, one of them repeated. As you can see, I keep the notion of repetition salient.) Presumably, he would have said that 'color' contains only five (occurrences of) letters. But Canadians spell the same word with a 'u', 'colour'. This worried me. How many letters *are* there in the word *color*?

As a frequent, though inadvertent, practitioner of deviant spellings, I felt confident that there were particular words of English that I was misspelling, just as there are particular words of English that Canadians mispronounce, for example, the word 'about'. Wait a minute! That's not a mispronunciation; that's a mere difference in dialect.

So it isn't wrong if enough people do it. (This is a lesson I'm trying to protect my grandchildren from learning.) The problem with my misspellings is that they are idiosyncratic—not enough people do it. (Here we see the tyranny of the majority in one of its ugliest forms. Why aren't the Libertarians doing something to protect me from the heavy hand of literary orthodoxy?) It is interesting how tolerant we are of deviant pronunciations and how intolerant we are of deviant spellings, especially since spelling, a relatively recent invention, is only a way of transcribing pronunciation. (I don't mean to whine, but this has been a heavy burden on me.)

All this led me to think that words could not be identified with either spellings, that is, with strings of letters, or pronunciations. Strings of letters and pronunciations are the shape-like features of words against which we measure perceived inscriptions and utterances in order to make out what words are presented. But shapes could not be what words *are*; they must be something more abstract, something that *has* spellings and pronunciations.

Another important motivation for me was that I was interested in how to think about contingently existing, seemingly abstract objects. These are objects that are natural, that were *created in nature*, by a (perhaps) human creator, at a time, in a place, and that live their lives in nature and can change over time—as we do, though we are not abstract objects—and that can, under certain natural conditions, cease to exist. Some artistic creations are material: sculptures, paintings, rugs, but some are seemingly abstract: stories, songs, ballets.[6] Where do these objects belong? The conception of a world of the eternal—unchanging and occupied by abstract objects—and a world of nature—in constant flux, occupied by material objects that *resemble* their abstract counterparts—seemed too disjoint, even given the fact that we have cognitive access to the abstract world. Words seemed the perfect vehicle to shatter this conception. Anyone who has traveled the British Isles quickly learns that what makes two utterances utterances of the same word is *not* that they resemble one another.

---

[5] W. V. Quine, *Mathematical Logic*, rev. ed. (Cambridge: Harvard, 1951), p. 24. According to the 2010 United States Census, the population of Boston is now about 600,000, though the great geographer's estimate is correct for 1950.

So that's the background.

Returning to Hawthorne and Lepore's paper, I want to say how much I welcome their very scholarly and thoughtful investigation of the topic I wanted to get people to think about. They know vastly more linguistics and bring vastly more scholarship to bear than I could ever muster. I learned a great deal from their paper, which is rich with erudition, argument, and counterargument. Their paper has the tone of disagreement, but when I think about what they are saying, I feel that for the most part it is quite supportive.

Here is an example from early in their paper:

> The first critical point we wish to emphasize is that we quite agree with Kaplan that a philosophically satisfying theory of words cannot proceed entirely within a shape- or form-theoretic framework.

I recognize that they regard it as a "critical" point, but somehow I don't feel criticized.[7] Be that as it may, I really enjoyed reading their paper, and rereading it, and I hope I can raise a few issues that may be fruitful for further discussion. My comments are confined mainly, but not entirely, to the few points of disagreement.

As they say in their introduction, their discussion divides into four parts. In Part One, they take up my argument against the *token-type* conception and in favor of what I, probably misleadingly, call the *stage-continuant* conception of what a word is. In Part Two, they discuss further elements and themes of my theory, including the role of repetition and my view about what they call *the constitutive authority*

---

[6] Russell seems to have regarded properties and relations as existing only contingently. He writes in his introduction to the second edition of *Principia Mathematica* (Cambridge, UK: University Press, 1927), "Logic does not know whether there are in fact *n*-adic relations (in intension); this is an empirical question." I might have called the objects I am interested in *empirical abstract objects*, but some of the characteristic features I find in my created objects do not seem to apply to properties and relations in general.

[7] Hawthorne and Lepore, *op. cit.*, p. 451. I note that the published version of the article replaces "critical point" with "point we wish to emphasize," thus cutting the ground out from under a perfectly good joke for nothing more than clarity.

of intentions. We have a real disagreement about the constitutive authority of intentions, but aside from that I am pretty much in agreement with what is said in these two parts. In Part Three, they either suggest that there are no such things as what I call *common currency names,* or else they claim that they are badly described and should be thought of merely as uses of what I call *generic names.* I am not very invested in the terminology, though I continue to think that mine is better for reasons that I will mention. When they say in their discussion of common currency names that my advocacy for such entities, however "thought of" and whatever called, "[raises] a suspicion that he may be in pursuit of a chimera," it sounds as if they mean to doubt the very existence of these things.[8] Finally, in Part Four, they discuss criteria for word identity. Put in terms of "criteria," this sounds technically challenging. But Frege's worry about the difference between 'A=A' and 'A=B' presupposes the distinction between using the same name twice and using two different names. This is not to say that there are no hard cases, only that it is an important distinction, with or without criteria.

I will respond to each of the four parts separately.

PART ONE: TOKEN-TYPE VERSUS STAGE-CONTINUANT

I thank Hawthorne and Lepore for documenting the prevalence of authors who take orthographic and/or phonetic shape or form as the criterion for word identity. They establish that I was not going after a straw man. (Whew!)

I was surprised that they read my stage-continuant terminology as "using a framework that is standard to four-dimensionalist metaphysics."[9] I am not exactly sure what that framework is, though I think it has to do with things being space-time worms, or perhaps regions of space-time, or perhaps fusions of parts—at any rate, things that do not change but only have temporal parts in the way those of us that do change (alas) have spatial parts. This was certainly not what I had in mind. In fact, my paper emphasizes the notion of change for words, especially changes in spelling and pronunciation. I think that because they know so much more metaphysics than I do, my terminology caused them to think that I am deeper than I am. But I am not.

My talk about a continuant was partly to affirm my creationism; words are worldly *created* objects, not denizens of a timeless eternal place. Talk of a continuant also helps to emphasize that the life of a word has *continuity* through time. There may be temporal gaps between its utterances, but if so, it has been stored somewhere, as a physical token in a manuscript or possibly in a mental lexicon. It is this continuity of earthly embodiment that makes two utterances utterances of the same word, not some form of phonological or orthographic resemblance to an ideal form. Related to my creationism is the fact that the words I was aiming at all along were names, in particular *given* names, like 'Hesperus', 'Phosphorous', 'Londres', 'London', 'David', and perhaps even 'Paderewski', though that is an inherited name, not a given name, properly so called (his given names were 'Ignacy' and 'Jan').

---

[8] These suspicions arise immediately following their initial, deflating "Nevertheless."

[9] *Ibid.*, p. 460.

I have now concluded that the token-type terminology is too powerful and too useful metaphorically to make it the focus of attack. I never meant to attack the abstract notion of a kind versus an instance of that kind. This is a useful idea, although there is an interesting literature on whether *types* should be thought of as *kinds*. What I wanted to attack was the idea that the type was an ideal, Platonic form living in an eternal, unchanging world and that what made a token a token of that type was that it *resembled* it. So what makes two utterances utterances of the *same* word is that they have the same phonetic form, or at least resemble one another closely enough in phonetic form, and what makes two inscriptions inscriptions of the same word is that they have the same orthographic form, or at least resemble one another closely enough in orthographic form. Sounds and shapes are kinds that have instances which must resemble one another in critical ways. I call your attention to the fact that Charles Sanders Pierce, who introduced the token-type terminology, also referred to tokens as *replicas*. So he seems to have had the resemblance model in mind. But words are not that kind of kind. Words are more like the Kaplan family. We are a disparate bunch; we don't look or sound much alike. But we are all members of a single family, a single kind, if you will. What connects us are certain relations, but they are historical in nature and not apparent to perception.

I have a very naturalistic conception of words. Words are earthly, created things. What makes two utterances utterances of the same word is that they descend from a common ancestor. This no more requires them to resemble or *replicate* that ancestor than my children are required to resemble or replicate their parents in order to be members of my family.

I think that Hawthorne and Lepore agree with me on this, my most important point. My reference to Pierce indicates that what I was attacking is not a straw man, and Hawthorne and Lepore powerfully document this fact in the opening pages of their essay. However, by the time we get to the end of their paper, they are worrying, with some justification, about my creationism. They have a number of worries, including whether a word need be created in a single earthly event. My basic idea was that a word, a certain lexical item, something like what they end up calling a *lexeme*, is created in the course of some event, for example, a name-giving. The word then makes its way in the world by being passed from person to person through utterances and inscriptions—*perceivable*, though not necessarily *material* entities—and when not being slung around, it is stored away in a *mental lexicon*, ready and waiting to be uttered or recognized in the utterance of another.[10] In the course of a word's travels through the world, its sound may change and its inscriptional shape may change, but its identity is maintained by its utterances and inscriptions being, as it were, nodes on a single, continuous tree of utterances, inscriptions, and quiescent storage. It is this sort of fanciful imagery that led me to think of individual utterances and inscriptions as stages in this metaphorical tree. The growing, ever changing tree itself I thought of as the continuant, that which maintains its identity throughout these changes.

However, as I said, I think I may have made some bad terminological choices. So let's return to the token-type terminology but this time throw away the associated ideology of ideal forms and resemblance. Here I endorse what Hawthorne and Lepore call their "second—more critical—point":

> One can favor the type-token conception in thinking of a word as an abstract object without taking on board any commitment to a shape- or form-theoretic conception of words.[11]

Reverting to this terminology also allows us to get a better grip on the notion of an *occurrence* of a word. In my view, word *types* are earthly created objects. In my original paper I put it this way:

> One might think of them as trees. Stemming out from their creation, with physical and mental segments; the mental segments able to produce many physical branches and able to merge many physical branches, and the physical segments each stemming from [their own] mental segment and able to produce many mental branches....they are objects of the created realm, created by language makers. The world is not brimming with unspoken words. Words never actually created *are* not.[12]

---

[10] On inscriptions being perceivable though nonmaterial, note that when we inscribe letters in stone—"It is written in stone!"—it is the space that we perceive, not the hunk of stone. For the latter, it would be necessary to out-scribe the letters.

[11] *Ibid.*, p. 452.

[12] Kaplan, *op. cit.*, p. 117

This harsh ontological banishment from the world of ideals does not apply to compounds built from word types and other linguistic types such as prefixes, and so on. These I take to be abstract structures containing words as constituents. My creationism about words does not extend to sentences. The world in which sentences and other compounds live *is* brimming with untokened types. Put roughly, the basic elements of the language are earthly creations, but the compounds generated by syntactical rules (the rules also being earthly creations and thus subject to change) are structures—*types*—which may or may not have tokens.

Thus the sentence type 'Help, help, my hair is on fire', which may have had its first token when I wrote my notes, contains two occurrences of the word type 'help'. Since I believe that when we talk about words and sentences it is mostly word types and sentence types that we refer to, I would just say that the sentence has two occurrences of the same word. My utterance, which was a token of the sentence, contained two tokens of the word. I would also be inclined to say that the token sentence contained two occurrences of the word, each occurrence being a distinct token. (I suppose we could use a single token for two occurrences of a word if a discourse were slowly scrolling across a screen and people behind the screen removed word tokens from one end and queued them up again at the other. This actually opens a rather complex subject that I have written about elsewhere but do not want to get into here.[13])

Back to Hawthorne and Lepore's first critical point, quoted above, that they agree that a philosophically satisfying theory of words cannot proceed entirely within a shape- or form-theoretic framework. They expand on the point in a learned and gratifying way. It turns out that the word 'color' has no fewer than 20 (!) different spellings, including 18 that are now out of fashion.[14] This was huge for me. I had worked hard to make this argument. If we agree on this, I feel that my main battle is won. Generally speaking, here included, Hawthorne and Lepore worked scrupulously to explore arguments counter to mine. That is part of the reason I was so gratified that they accepted this claim as manifestly correct. (There are other arguments for a resemblance-based theory that they did not explore; for example, what makes it possible to recognize new tokens of a word that one already has in one's lexicon?)

---

[13] Kaplan, "Reichenbach's *Grundzuge der symbolischen Logik*," in Hans Reichenbach, *Gesammelte Werke in 9 Bänden. Band 6: Grundzuge der symbolischen Logik*, ed. Maria Reichenbach and Andreas Kamlah (Wiesbaden, Germany: Friedr. Vieweg & Sohn, 1999), pp. xi–xxii.

[14] Among the literary orthodoxy.

Their second critical point—that the use of the token-type language does not commit one to a shape- or form-theoretic conception of words—leads to an extended discussion of what they take to be my four-dimensionalism. As indicated above, that is nothing like what I intended. (Again, I am not contending that the reading is unfair. I can sort of see how one who is versed in deep metaphysics might have come to think that. On the other hand, there are explicit counter-indications in my paper, including the fact that the entire final section is devoted to the modal properties of words. Hawthorne and Lepore "anticipate that insofar as one pursues a continuant model of words, one will likely be drawn into [a counterpart-theoretic] account of attributing modal ascriptions to words."[15] But I was not. I have long forsworn counterpart theory.[16])

They contrast my *stage-continuant model* with what they called the *abstracta-articulations model*, which treats "a word as a nonconcrete object which is articulated by various concrete events or objects."[17] I am not absolutely sure how to read this, but it sounds plausible to me. You can see from my talk about trees that I am a little ambivalent about whether to regard words as abstract objects at all. They are not material objects; not even their tokens are necessarily material objects, as the cases of words incised into stone and stencils show. So maybe *abstracta* is good. I take it that in gracious deference to my token-type phobia, they use 'articulation' rather than 'token'.

But there is one thing that this model seems to miss: my concern about *continuity* for the existence of a word. Russell once said that if no one thought about Hamlet, that would be the end of him. (He was contrasting Hamlet with Napoleon, who would have soon seen to it that someone *did* think about him.)[18] I think Russell is right about this (if we expand the cessation of thought about Hamlet to the destruction of all written and recorded references to Hamlet, and we understand *ceasing to think about* as forgetting in the sense of losing the *power* to think about Hamlet). I believe that Russell's picture of what maintains such an object's existence, along with some version of my creationism, is characteristic of the kind of worldly abstract entity I was trying to write about.[19] The

---

[15] Hawthorne and Lepore, *op. cit.*, p. 454.

[16] Forsworn shortly after writing "Transworld Heir Lines" in 1967 The paper later appeared in Michael J Loux, ed., *The Possible and the Actual: Readings in the Metaphysics of Modality* (Ithaca: Cornell, 1979), pp. 88–109.

[17] Hawthorne and Lepore, *op. cit.*, p. 453.

[18] In the "Descriptions" chapter of Russell, *Introduction to Mathematical Philosophy* (New York: Macmillan, 1919).

[19] I would not adopt the same view about numbers. If everyone ceased to think and talk about prime numbers, that would not be the end of them. They would wait—passively (unlike Napoleon)—to be thought about and talked about again.

"end" of Hamlet comes about when there can no longer be a continuous path from Hamlet's creation to a new thought about him or a new use of his name. If, after the end of Hamlet in Russell's sense, a million monkeys working at typewriters should produce a play that is orthographically identical to the play *Hamlet*, I would regard it as a *new* work of fiction ("Created by monkeys!") containing a new common currency name for a new fictional prince. In the realm of created objects, look to continuity, not resemblance, to determine identity.

Continuity is a part of my naturalism. When mankind becomes extinct and our records turn to dust, the words will go with us. They will not remain in some prolix realm waiting for a new species to perform them.

Do things like Hamlet and 'Hamlet' have a continuous existence?[20] We do continue to think and talk about Hamlet, but not continuously. We pause, sleep, eat, talk about other things, and perhaps later return to Hamlet. How do we manage to return to Hamlet after a gap? Storage! This is why I insist on including the quiescent periods of storage of a word—inscribed, recorded, or stowed away in a mental lexicon—as among the stages of the word. There are performances of a word (events when a word is *in performance*, when it is on the move): utterings, auditions, inscribings, readings, and then there are quiescent stages of a word as it lies unread in a dusty manuscript in a dark archive. If the word 'Hamlet' were "ended" by our not only ceasing to perform it but by destroying these stages of storage, I do not see how that very word could be rediscovered. The "end" of 'Hamlet' comes about when there can no longer be a continuous path from the name's creation to a performance of it. And I think the same for Hamlet. In fact, I also think the same about Napoleon. Note that by the time Russell wrote, if we ceased thinking and talking about Napoleon, he no longer had the power to see to it that we began again. But we have not ceased thinking and talking about him; we still have the ability to do so. (And not "by description," as Russell would have it, but directly.) Again, I think that a continuous path from Napoleon to our thought is required.[21] In my old paper "Quantifying In," I thought that this path involved a causal connection, being *en rapport*. But in "Words," I abandoned the view that the continuous path I insisted upon for word identity required causality. This was because the continuity involves mental activities that I am methodologically reluctant to characterize as causally determined. The continuity comes about (in part) from links between storage in a mental lexicon and utterances that draw upon that very stored lexical item. The question of whether the words we store and the words we utter are causally determined or freely chosen seemed irrelevant to the theory of words I aimed to promote. The important thing is that there *be* a link, not that it is causally determined. So, a chain? Yes. A continuous chain with active and inactive links? Yes. A causal chain? Who knows!

---

[20] Words (and not just ambiguous words) are Hamlet-like. Both have physical embodiments from time to time, and both are dependent on thought and memory (including external memory) for their continued existence.

[21] As an historical aside, Russell's actual point was that Napoleon had a robust form of reality, like that of the moon, entirely independent of thought and talk about him, whereas Hamlet would never have had *any* form of reality, not even an enfeebled one, had there never been thought or talk. By this measure, the word 'Hamlet' has no more reality than Hamlet; nor does any other word. But I doubt that Russell would have been happy with this result.

This continuity feature of the worldliness of words is an important part of my ambivalence about whether to say that words are abstract objects. However, my *continuant*, the tree of what I unfortunately called "stages," is itself probably best thought of as an abstract object, a structure, a network, whose constituents are utterances, inscriptions, and so on, and what I called the "more mysterious intrapersonal stages"[22] (by which I was referring to storage in the mental lexicon). There should be agreement among us that words have modal properties. Certainly, particular words of English might never have been created, might have had a wider distribution earlier in time, might have gone through dialectal changes in ways other than they did, and so on. And some have worried whether abstract structures in and of themselves, that is, not "under a description," have modal properties.[23] So one way to think of words is as abstract objects that are sui generis but *associated* with trees of their tokens (if it is correct to describe a word's mental stages as "tokens," which seems questionable), where the tree satisfies the continuity condition, and the words, though they are abstract objects, live and die with their associated trees. In an interesting footnote, Hawthorne and Lepore attribute to Mark Johnston the suggestion that "words (and species) are higher-order persisting objects that are concrete; that are to be sharply distinguished from their realizations in utterances, inscriptions, and so on; and that...are best thought of within the framework of three-dimensionalism."[24] That sounds a lot like my stage-continuant model, in which the continuant is what persists and the "concrete" part accounts for their living and dying with the associated tree. Maybe that's what my nuclear family is, also. I like the idea that my nuclear family is a persisting object and that it is of a higher order. In the end, as Hawthorne and Lepore claim, my ambivalence about calling words abstract objects is probably due to a poorly thought-through metaphysics. And I am not confident that I have done better here.

---

[22] Kaplan, "Words," p. 98

[23] To my ear, it is acceptable to say that a tree, whether abstract or material, might have branched earlier.

[24] Hawthorne and Lepore, *op. cit.*, p. 453n20.

This leaves open the question of whether Hawthorne and Lepore's abstracta-articulations model takes account of storage in the mental lexicon. It probably does, because they speak about mental symbols in the language of thought. So perhaps they regard these mental symbols as "articulations" of the word. Finally, returning to my main point about continuity, my use of 'continuant' emphasizes the importance of continuity to the existence of this whole realm of worldly, created, abstract objects, and I am not sure that the abstracta-articulations model captures this requirement.

I believe I have already responded to Hawthorne and Lepore's second concern. It is my claim that "the world is not brimming with unspoken words."[25] They cite compound words such as 'unhappy' as a counterinstance. They anticipate my reply that compounds are not earthly creations, and even seem to acknowledge that the restriction of my view to the kind of linguistic entity to which it was intended to apply will handle all of their concerns but one.[26] Their unanswered concern is about words introduced descriptively. Since, as noted, my primary concern was with given names, let me use a variant of their example, one that involves such a name. Suppose I say to my wife, "Let's call our first child, if it is a boy, by the name whose first syllable is the first syllable of *my* paternal grandfather's paternal grandfather's name and whose second syllable is the first syllable of *your* paternal grandfather's paternal grandfather's name, and, if it is a girl, by the name whose first syllable is the first syllable of *your* maternal grandmother's maternal grandmother's name and whose second syllable is the first syllable of *my* maternal grandmother's maternal grandmother's name." Now suppose that our firstborn arrives, but before we get around to using this name, let alone figuring out what the name actually is, the world blows up. Hawthorne and Lepore claim, "It is intuitively clear that [such] speech serves to *introduce* [a word] into the language."[27] Does it? Is it *intuitively clear*? To be accurate, although they speak of introducing words "descriptively," their examples are all such that a token of the word immediately pops into one's mind, examples like: let the first syllable be 'un' and the second syllable be 'voke'.

---

[25] Kaplan, "Words," p. 117.

[26] Hawthorne and Lepore, *op. cit.*, p 457, second full paragraph

[27] *Ibid.*, p. 456

Once we store 'unvoke' in our mental lexicon, that is good enough for me, even if we never utter or write it. So I think I can account for their cases, and I think that my own case—a paradigm of what Russell once called "knowledge by description without acquaintance"—is such that it is just hard to decide whether it succeeded in introducing a new, but never used, word into the language or not. I am inclined to think not.

ON TO PART TWO

In Part Two Hawthorne and Lepore take "a number of themes that shape" my discussion and adopt them as "framing principles" for their own discussion. But they also take up two of my central ideas that strike them as "seriously problematic."

The first of my themes that they adopt is what they call the *Evolutionary Constraint*. (As an aside, I would not call these *constraints*; I see them as ridding us from the yoke of *formal resemblance*.) Here is the evolutionary constraint:

> One of Kaplan's guiding insights concerns the evolution of words: over time there may be considerable variation in how a word is written or pronounced. As it is transmitted from one epoch to another, it can change not only its phonological and orthographic contours, but its semantic and grammatical ones as well.[28]

As noted, I am delighted that we agree on this. In "Words," my own thinking was more along the nature of idiolectal variation due to damage to, or a disability of, a person's auditory or articulatory systems, rather than the kind of dialectal variation that they seem to have in mind, but we will come to that.

The second theme that they adopt is what they call the *Multiplicity Constraint*:

> [A] word can be articulated in untold different systems. One and the same word can be written on a pad with a pen, typed on a sheet of paper, projected on a screen, spoken out loud, signed with a gesture, and Brailled on a plaque. Different media can be used to express the same message. Only our imaginations limit how we go about articulating words.[29]

I trust that these "articulations" include presence in a mental lexicon, since that notion is fundamental to my conception. The word 'articulation' can refer to an act or to what is produced by the act, to an uttering or to the utterance, to an inscribing or to the inscription.[30]

---

[28] *Ibid.*, p. 459.

[29] *Ibid.*

[30] Later on the same page, they speak of a *performance* of a word, which is subject to the same ambiguity. Does the written word, stored in a dusty manuscript in a vault, continue to be a performance?

I did not use 'articulation' in my original paper; instead, I spoke of *tokens*, and then, after my animadversions on type-token theories of words, I called them *stages*. I was speaking of the product of the act, not the action. This is important to me since my continuity claim for the existence of a word would plainly be false if the stages consisted of acts of production and reception. In that case there would be gaps in the life of a word as it lay quiescent in a manuscript or a mental lexicon. The fact that Hawthorne and Lepore speak of the "archipelago" of the utterances of a word and describe my notion of a continuant as the "fusion" of such utterances and inscriptions causes a worry that they have not caught my drift, since an archipelago is, by definition, discontinuous.

There are some interestingly hard cases in the category of articulations. It seems natural to include a score as an articulation (token) of a song, since many people can sight-read a score and hear the song. I am more worried about words recorded in digital media, as they are in our computer storage devices and in digital audio recorders. I do not know anyone who can sight-read a DVD. But perhaps the fact that we need transcriptional assistance from an artifact should not be decisive. After all, we need lemon juice and heat or ultraviolet light to write and read invisible writing, and who would deny that invisible writing contains tokens of words?

The third theme that they adopt, and one that is of great importance to me, they call the *Coincidence Constraint*.

> [T]wo utterances or inscriptions, even within a particular community, may in an important sense of the word 'word' count as utterances or inscriptions of different words even though they are exactly alike in how they sound (in the case of two utterances) or in how they are written (in the case of inscriptions).[31]

Unfortunately, they resist one of my central claims, that my first name and Hume's first name are different words. In fact, they seem to resist the claim that my first name and Hume's first name are importantly different lexical items of any kind. I will return to this when I discuss their Part Three.

The first of the two themes they criticize in Part Two involves a claim that I do not think of myself as having made. It is that when one utters a word that one is not in the process of creating, it should be conceptualized as a case of repetition. They interpret this as if I were saying that every time one uttered a word, one had in mind the source from which one first heard that word. It's hard for them

CENTRAL LIBR. UNIVERSITY OF CALCUTTA

---

[31] *Ibid.*

believe that I said this; it is so manifestly false. I don't recall ever thinking it. If I did say it, I didn't mean it. Words that we haven't created we learn from someone else. So my picture is: we hear (or read) the word from someone else, store it in our mental lexicon, and on a later occasion draw it from the lexicon and use it (or, use the mental lexicon to recognize another instance of that same word). This is fundamental to characterizing the tree whereby words are transmitted. One *could* say, very roughly, that the word uttered repeats the word as first heard and stored, but this would be without any suggestion that storage of the lexical item includes storage of the details of the event of first hearing the word. We may see and/or read multiple instances of the word before using it.

The second theme they reject also involves the notion of repetition and my claim that a sincere subject, intending to repeat a word that has been uttered by an examiner, will, indeed, utter that word. Here is what I say (including some of the surrounding material of the passages they cite).

> [I]ndividual differences in the physiological processing at these psycho-physical transition points [the ear and the organs that produce sound: the mouth, tongue, lips, vocal cords, and so on] may make it that what comes out is not going to *resemble* what went in. However, and this is what is important for this thought experiment, the exact functioning of the psycho-physical transition mechanisms is irrelevant to our characterization, our intentional characterization, of the case as one of '*repeating* the word he heard'. He may not do it well, from the external point of view. But it *is* what he is doing. We don't question that the connection has been made in the black box. No matter how poor the subject's *imitative* ability (his ability to make his output resemble his input), we can imagine circumstances in which we would say, "Yes, he is repeating that name; he is saying it in the best way that he can."[32]

> [I]t is beyond doubt that the utterance or inscription transmitted *could* be an utterance or inscription of the same word as that received, although...the difference in sound or shape or spelling, can be just about as great as you would like it to be.[33]

Against these claims they make two interesting counterclaims, one a needed correction that I wish I had thought of and the other a claim I will dispute.

The needed correction is this: suppose someone has had a terrible accident. We ask him for his name and telephone number. He has the name and telephone number in mind and strives to speak it; he

---

[32] Kaplan, "Words," pp 103–04.

[33] *Ibid.*, p. 101.

intends to be uttering his name and telephone number. But all that comes out is a monosyllabic grunt that sounds a bit like the word 'row'. Does my view that intention makes it so imply that the grunt is, in fact, an utterance of the name and telephone number? It need not, because we can take advantage of Hawthorne and Lepore's excellent suggestion that the right thing to say in this case is that the injured person cannot speak. He didn't say 'row', and he didn't say what he was intending to say, namely, his name and telephone number, he didn't say anything at all. *He cannot speak.* Suppose your dentist injects so much Novocain into your mouth that you can only produce inarticulate grunts when you attempt to ask me to "Please pass the beets." Must I say "Yes, you are saying 'Please pass the beets'; you are saying it in the best way that you can"? No! The more plausible response may be Hawthorne and Lepore's, you didn't say "Please pass the beets." You didn't say anything at all. You cannot speak.

The point I wish to dispute is their doctrine in connection with mispronunciation that *toleration has its limits.* (Shouldn't any doctrine so called be regarded with suspicion?) They elaborate this idea by referring to the standards of the *local* community. Then they correct this by saying it is the standards of the *relevant* community, citing the case of someone who gives a performance in which she attempts to simulate a fourteenth-century English speaker, an exotic case in which widespread intelligibility is not the primary aim.

> *Tolerance*: Performance p is of a word w only if [p] meets relevant [local] performance standards.[34]

Aside from the *1984*-style use of the word 'tolerance', this claim seems to me just to ignore a raft of counterinstances. The doctrine asserts that if the performance standards of the local community have not been met, no words have been said.

Suppose, as they agree can happen, pronunciation evolves in radically different ways in two isolated communities. We do not question that the same words are being said in both communities, but dialectal differences are very great. Then globalization strikes, and a member of one community wanders into the other. He will not meet local standards, if meeting local standards means being locally intelligible.[35] It is communication with the local community that is his primary aim; he is not aiming to show off his exotic dialect. Surely he is saying words. Or consider the case of the person with extremely illegible handwriting. Everyone agrees that the note Philippa Foot sent me was filled with written words, words of English, words in our dialect of English. But few of these words are recognizable. According to Hawthorne and Lepore, this familiar situation is impossible, a contradiction. In these cases, their clever way out in the case of the accident victim or the Novocain-loaded speaker is not available. We would not say that Foot was not capable of handwriting. She wrote a lot. She just wrote so idiosyncratically that most of what she handwrote is unintelligible. In all of these cases, I stick with my view of the authority of the writer. She knew what words she was writing at the time, and she did write those words. And she was writing words of English—British English, but English all the same. She was not writing in a foreign language or using a secret cipher. Perhaps she could always read her own writing (I can't always read mine), but of course she wrote many messages to other people. It seems fantastic to me for those other people, the people of the local community, to insist that because they cannot *read* the words she wrote, she did not actually *write* any words. Talk about the tyranny of the majority! Philippa might have apologized for her bad handwriting, but she certainly should not have apologized for failing to include words in her letters.

---

[34] Hawthorne and Lepore, *op. cit.*, p 463.

[35] As globalization increases and a sufficient number of members of the traveler's community arrive, the locals will attune their hearing to the traveler's dialect and it will become locally intelligible But this need not happen immediately.

Hawthorne and Lepore directly address the case of the Reverend Spooner, who is unreliably reported to have said, "God is a shoving leopard." This is a seemingly intelligible, if blasphemous, claim by the good Reverend. But of course those were not the words he was saying. The words he was saying were, "God is a loving shepherd," but because of his eponymous speech defect, they came out badly. Hawthorne and Lepore do not recognize this as a bad performance; they claim that he has said nothing at all. My goodness! The guy has a speech disability. Show a little tolerance.

It is one thing to say of the injured person or the person full of Novocain, speakers whose own performance standards match those of the community, that they are unable to speak. They might well agree. But it is quite another to say of someone with a speech disability, a strong foreign accent, or an unfamiliar dialect that *they* are not speaking words. They would never agree. This is an alternative to Hawthorne and Lepore's "prevailing community standards" test for failure to speak words. Ask what the speaker would say of his or her own performance.

One might think that it makes no difference whether we say that Philippa *wrote* such-and-such words or that Philippa wrote nothing, though she *intended to write* such-and-such words, whether we say that the Reverend Spooner *said* that God is a loving shepherd or that he said nothing, though he *intended to say* that God is a loving shepherd.

But the line between *saying* and *intending to say* should not be drawn carelessly. A doctrine that encourages us to apply prevailing community standards before examining intention is bad social policy. It is stultifying. It self-righteously shifts the blame for failures of communication to the outsider and reduces *our* responsibility to attune ourselves to nonstandard speech. I see it as an evil form of nativism. If philosophical theories could be dangerous, I would worry about this one. (When they say that their principle "is anathema to Kaplan,"[56] they certainly hit the nail on the head.)

Now, I know John and Ernie well enough to know that if a nonnative speaker or a speaker with a speech disability were to ask them for help, they would patiently struggle to understand his words so long as they could tell that he was attempting to say words of English. They certainly would not abandon the attempt to understand him simply because he was not meeting prevailing community standards.

So I think that Hawthorne and Lepore are wrong here in a fundamental way. Once we pass the very low threshold of whether the person is able to speak or write at all,[57] that is, *whether* words are used, the standard for *which* words are used is not *local intelligibility* but *speaker intention*.

PART THREE

In Part Three, Hawthorne and Lepore express skepticism about my notion of a *common currency name*. I hardly know what to say about their objections, which seem to me to range far afield. (Perhaps this is the price of erudition—or my lack of it.) I basically recognize two kinds of names: *generic names* and names given to particular individuals. The former, generic names, are those found in the lists of names in books titled *What to Name the Baby* and in the Social Security Administration's list of the most popular names given each year. The latter are singular terms, *proper* nouns; each refers to (or purports to refer to) a particular individual. In paradigmatic cases they refer to the individual to whom they were given. The generic names are not singular terms; they do not name anything. Because it is the latter kind of name, the singular terms, that we *use*, I adopt a phrase of Kripke's and call them "common currency" names. Mark Johnston might describe the generic names as *higher-order* names, a word type of word types. When I say "John and Ernie wrote a valuable paper," I am using the utterances 'John' and 'Ernie' as singular terms, to refer to particular individuals, presumably in virtue of their having been given these names, probably by their parents or, in the case of 'Ernie', having been given a name for which this name is a standardized brief form (a topic I take up at some length in my original paper).

---

[56] *Ibid.*

[57] I acknowledge that this threshold is vague, but would urge that it not be measured against prevailing community standards of intelligibility but rather against something closer to the speaker's own standards. Judgments of nonperformance (as contrasted with bad performance) should presuppose some form of buy-in from the speaker.

The names we *use* are the singular terms, what I call *common currency* names. (This is why I wrote, and they quote me on this, "for serious semantics...it is my common currency conception that would be important."[38]) We do not *use* generic names, because they do not name anything. We talk *about* generic names. Hume and I have the same first name. Our common first name is the *generic* name 'David'. Since common currency names are singular terms and Hume's first name names him and my first name names me, they could not be the same *singular* term. So I conclude that Hume and I share a generic name but have distinct common currency names.

We could create a little puzzle here. Hume and I have the same first name. But Hume's first name names him and not me, whereas my first name names me and not him. How could this be? Call Leibniz! Paradox! The solution *two kinds of names* suggests itself. The analogy to the type-token distinction, used metaphorically as of the relation of kind to instance, is almost irresistible.

My creationist, worldly metaphysics of words also brings to the fore the many nonsemantic differences between Hume's common currency name and mine. His common currency name is in the lexicon of a vast number of people worldwide; mine is in the lexicon of a relatively tiny number of people mostly located in Southern California. My common currency name was created in an event around the time of my birth, when I was named, and Hume's common currency name was created much, much earlier, when he was named. His came into existence earlier than mine and will no doubt live longer than mine. There is every reason to think of these two names as different words. Our common generic name has a very long history stretching back to biblical times and beyond (if there is a beyond). It was evolving in pronunciation and spreading long before either Hume's or my common currency names were created.

Many of these ideas seem so commonsensical to me that it is hard for me to get a grip on the objections.

Here is a thought experiment. Suppose that our naming practices did not allow us to reuse an old name—no naming one person in honor of another, and so on. Each thing that was given a name had to be given a name distinct from all other names, the way in which the Social Security Administration tries to give each of us a distinct numerical index. What would we say about these names, about their creation and their use, about their semantics? Wouldn't they be exactly my common currency names? Generic names are exotic birds; we have them to thank for the Paderewski puzzle.[39] But the real action, in thought and in speech, is with the common currency names.

---

[38] Kaplan, "Words," p 111.

At one point, if I understand them correctly, Hawthorne and Lepore suggest that we can "eliminate the need for an ontology of common currency names" by speaking instead of *uses* of a (generic) name.[40] This is the sense of the word 'use' in which one might say there are two uses of the word 'bank'. So there is one *use* of the generic name 'David' to refer to Hume, and there is another *use* of the generic name 'David' to refer to me. These *uses* of the generic name are singular terms. The singular terms that we all use are not common currency names—that ontology has been eliminated—they are *uses of generic names*. But don't kid yourself that we have eliminated an ontology. We have simply traded an ontology of common currency names for an ontology of *uses* of the generic names. The same number of items will occupy each ontology. The same principles of individuation will apply to each ontology.

The new terminology seems awkward to me. Instead of saying, as I do, that we use names to refer to individuals, we must now say that we use uses of names to refer to individuals. 'Hesperus is Hesperus' involves two uses of one use of one name; whereas 'Hesperus is Phosphorus' involves one use of one use of one name and another use of another use of another name. Names do not, as we have been taught, name anything. It is only uses of names that name things. Isn't it natural to regard these uses of names as *words* in their own right? Aren't the different uses of the generic name 'David' naturally thought of as different words? At any rate, I just do not see how anything is gained by the switch in terminology. Their argument for the terminological switch has the dismissive quality of *mereness* toward such uses, an attitude that is reflected in their claim to eliminate the need for an ontology. However, as I have argued, it is the common currency names that are central for semantics and the generic names that are merely unused abstractions. But for those who are more comfortable with the terminological switch, be my guest.

Here is a query based on the hypothesis that there is more than a mere terminological difference involved here. I believe that we store names of distinct individuals as distinct lexical items. This accords with my notion of a common currency name. If it is preferable to conceive of these items as uses of generic names, then it is only the generic names that should be stored in the lexicon, and there should be distinct pointers from each of the mental items that represent the individuals to a generic name in the lexicon. This is close enough to an empirical question that a really clever cognitive scientist ought to be able to find a way to run a test that will provide evidence for one view or the other.

---

[39] Saul Kripke, "A Puzzle about Belief," in Avishai Margalit, ed., *Meaning and Use* (Boston: D. Reidel, 1979), pp. 239–83.

[40] Hawthorne and Lepore, *op. cit.*, p. 470.

I am no cognitive scientist, but let me tell you a story about my mother's two Doctor Shapiros. (I know, 'Doctor Shapiro' is not a given name, but let it pass—it's an actual case.) My mother's primary care physician was Dr. Shapiro. He referred her to a specialist, another Dr. Shapiro as it happened. My mother reported her gratitude to Dr. Shapiro for sending her to Dr. Shapiro and compared Dr. Shapiro's virtues to those of Dr. Shapiro in a blithe piece of discourse, clearly oblivious to the homonymy. I was racing to keep up (which I was mostly able to do). But from her point of view, she was quite properly using two different words to refer to two different people. Why should there be a problem?

I have always taken it that the unselfconscious way in which my mother spoke of the two Doctor Shapiros was evidence in favor of the idea that there were two items stored in her lexicon. But I am no cognitive scientist.

It seems to me that a Hawthorne and Lepore-style theory of *uses* finds its proper home not in the realm of *names*, but rather in the realm of *indexicals and demonstratives*. There Hawthorne and Lepore are right: one word, many uses.[41] But as I have argued elsewhere, proper names are not indexicals.[42]

For completeness, I should also mention what I myself regard as *uses* of generic names: *common* nouns[43] (as opposed to the *proper* nouns that Hawthorne and Lepore regard as *uses* of generic names). This is the grammatical form that takes determiners and plurals. "Is she the Alice who sits in the front row?" "There are three Alices in the class." "There are many people named 'Alice Rivlin', but she is *the* Alice Rivlin." (To be *the* Alice Rivlin is to be the first person listed when you google the name 'Alice Rivlin'. Our co-symposiast is *the* Sylvain Bromberger.)

---

[41] One might, by analogy to proper names, claim that there are generic words: 'that', 'today', and so on, plus an unlimited number of distinct "common currency" words for each time the demonstrative or indexical is used. I see the analogy: only the "common currency" words actually refer, but I am not even slightly drawn to such an analysis.

[42] Kaplan, "Demonstratives," in Joseph Almog, John Perry, and Howard Wettstein, eds., *Themes from Kaplan* (New York: Oxford, 1989), pp 481–563.

[43] Mill claimed that the proper name 'God' is simply a monotheistic use of the common noun 'god'.

At one point, Hawthorne and Lepore suggest that context resolves the question of which *use* of a generic name is in play. If they mean to include speaker intentions in contexts, then I agree. (Though I do not think they do.) But if they mean to include only what is available to a standard auditor, what is common knowledge, then there seem to be obvious counterinstances. I did pretty well with my mother's two Doctor Shapiros, but not perfectly. And I was a *much* better than standard auditor of *her* speech and loaded with knowledge of the context. When I once said, "While on vacation, I found it difficult to swim to the bank," everyone assumed they knew what use I was making of the ambiguous word 'bank'. They did not know that in Hawaii they have poolside banking.

In the course of Hawthorne and Lepore's discussion of common currency names, they take up Kripke's Paderewski case. This is quite an interesting problem. I do mention it in "Words," but I do not think it will ultimately tell us whether there are common currency words or only uses of generic words.

Part Three is long and I have more to say about it, but this is not the place.

AND NOW TO PART FOUR

This section focuses on the criteria of identity for words, and more exactly, on the question of when two tokens are tokens of the same word. I wasn't really after exact criteria of identity. (I have some of the same skepticism that Hawthorne and Lepore express about such criteria. We haven't solved the problem for persons; why should words be that much easier?) I was after the *big picture* of what words *are*, and in particular, how the nature of *names* played into Frege's puzzle from the point of view of a direct reference semantics. Frege begins by assuming that 'A' and 'B' are different names, and arguing that if 'A' and 'B' had only *Bedeutung* and no *Sinn*, there would be no cognitive difference between 'A=A' and 'A=B', when 'A=B' is true.[44] As noted earlier, this seemed wrong to me, semantics aside, because if 'A' and 'B' *are* different names, the (linguistic) representations are different even if what they represent is the same, and *this* difference, the difference or sameness in words, is certainly a part of what is present in cognition.[45] It was this idea that I was trying to work out in my first two Ryle lectures.

In considering criteria of identity, Hawthorne and Lepore reject looking to origins (obviously *my* favorite in view of my creationism) because they think they have established the fact that there are unarticulated words.[46] First, this is far from clear to me. It presumably has to do with words introduced by description and never "performed," that is, words with no tokens. I responded to the case of words introduced by description in my discussion of Part One. Second, even if there are such words they would have origins, creators, and creating events (as I would put it). Third, even if there were unperformed, uncreated words that slithered into the language all on their own or propelled by systemic forces independent of any creator (remember we are talking about lexemes here, not compounds), wouldn't it be methodologically reasonable to regard them as exceptions, or belonging to a special category to be accounted for by an epicycle on the theory?[47]

---

[44] Gottlob Frege, "Über Sinn und Bedeutung," *Zeitschrift für Philosophie und philosophische Kritik*, c (1892) 25–50, translated almost everywhere.

[45] It is this claim that is challenged by Kripke's Paderewski cases.

In their discussion of a criterion they call *Connection*,[48] Hawthorne and Lepore seem to have dropped commitment to lexemes and reverted to compounds. The same compound could obviously have many first tokens. It is not a created word.[49] However, they do accept a principle, *Isolation*,[50] that I see (perhaps wrongly) as tightly connected to my creationism and its attendant continuity requirement, namely, that the same word cannot occur in causally isolated linguistic communities. This is a big step forward that I welcome. It tells us that even if the rivers and lakes of Twin Earth were filled with H_2O, their word 'water' would not be our word 'water'. Wow! But good!

They analogize words to songs and dances, which I was thinking about, too, when I wrote "Words." This, obviously, I like. More *created* things. They worry about evolutionary vagueness: when is a dance "inspired by" another but new, and when is it a mere variation? This worry applies also to generic names.

The notion of a variation on a theme is quite familiar in music. Many songs have the same chord progression. Almost all blues melodies do. So when a jazz player improvises on a theme, what makes it a variation on one song rather than another? I take it that the natural answer is intentionalist: it is the song that the player has in mind and takes himself to be improvising on. I am not sure whether to be somewhat less intentionalist about songs and dances and games than about words, but I am still pretty intentionalist. In a footnote they comment, "It is sometimes said that all that holds various games from different locales and epochs together is their name."[51] Probably they mean this just as a joke, but the use of the same name shows that the innovators *intended* the changes they introduced to be *variations* on the named game.

---

[46] They also implicitly reject the idea that two words can be performed at once. But this actually happens in puns.

[47] This is how I regarded answering machine messages and aphorisms in "Demonstratives."

[48] Hawthorne and Lepore, *op. cit.*, p. 478

[49] Perhaps they are thinking about words like 'smog' and 'brunch', compounds used to describe compounds. I suspect that these words retain a compound meaning in the following sense. If we were to discover that what we had called 'smog' never did contain smoke and fog, I think we would not conclude that we had discovered that smog contains neither smoke nor fog, rather, we would conclude that what we have called 'smog' was not smog.

[50] *Ibid.*

Whenever we look up the etymology of a word, including that of a generic name, we are faced with evolutionary vagueness.[52] However, although the elements of form—pronunciation and spelling—may change for a common currency name as it travels through different linguistic communities, we do not think of common currency names as having an etymology. The strongest case for my creationism is in connection with common currency names, and there we see variation (a child named in honor of a relative with a related generic name) in comfortable accord with the creation of something absolutely new (a name for this child).

Hawthorne and Lepore conclude on a metaphysical note: that there are natural-kind optimists—looking for joints, hoping for future research to help; and then there are gruesome pessimists—proclaiming that we live in a world of vagueness and uncertainty. These are large methodological issues. Interesting ones, but ones I do not wish to dip into (especially since I am an optimist residing happily in a world of indecision). But I wish to remind the pessimist that there is more to do than find exact criteria of identity for the elements of our ontologies. We can still correct misconceptions and make illuminating remarks about what a word is. A correct picture of what a word is can resolve many questions about word identities even if it does not resolve identity questions in every hard case.[53] The question of whether Philippa Foot's unintelligible handwriting contained words bears directly on the question of what tokens are tokens of the same word, as does the rejection of spelling as a criterion. Hawthorne and Lepore mark agreement with this sentiment when they say, "The good news is that the elusiveness of questions of word individuation need not indict our practice of positing words."[64] Good. We also posit persons even though there are hard cases, or at least hard *possible* cases, there too.

---

[51] *Ibid.*, p. 480n68

[52] I recently heard a noted evolutionary biologist explain on NPR that though parents and their immediate offspring always belong to the same species, over a longer period of time, ancestors and their ultimate descendents may belong to different species. He was calmer in the face of this astounding fact than Eubulides had been 2,500 years ago.

[53] Recall Kripke in *Naming and Necessity* (Cambridge: Harvard, 1980), when he says that although he is not confident that he can resolve the problems that the Frege-Russell theory of names is thought to resolve, he is confident that he can show that the Frege-Russell theory of names is incorrect. We make progress bit by bit

After giving us the good news, they give us the bad: "The bad news is that the accessible facts about words run so shallow that there is little philosophical payoff to ruminations about word identity." Uh oh, I think it's me, a notorious ruminator, that they are talking about. They continue, "Those who pursue questions of word individuation and hope for systematic answers are almost invariably in the grip of a faulty picture of the semantic mechanisms that underlie thought and talk about words."[65] This is a bold statement that hurt my feelings. It also confused me, coming as it does at the end of almost 40 pages of rumination about words and their individuation.

At the end, they despair of finding a criterion of word matching between performances. "Having eschewed superficial criteria, this challenge takes on a forbidding character, since there is no obvious surrogate that can provide a criterion once the form-theoretic ones have been dispensed with."[66] It was with pleasure that I wiped the blood of the form-theoretic criteria of word matching off my incisive arguments.[67] And they close with a sentiment that I guardedly endorse: "Rather than expect that such criteria will be forthcoming, we must take seriously a conception of our practice which guardedly endorses an ontology of words while despairing of such criteria."[68] My caution stems from my belief that many of their concerns about hard cases are more manageable when we limit ourselves to common currency names.

I have a final concern that the sort of criteria that Hawthorne and Lepore examine may not be the most important ones. In investigating word identity, I would focus more attention on the links in the *passage* of a word. The issue of whether first performances are the same seems to me less important than the issue of what makes the links hold. What makes my utterance of a word an utterance of the word I first heard from you? Here is where my example involving repetition is meant to provide an intuitive example of an output properly linked to an input. That link arises from perception and storage in the lexicon and later drawing from that lexicon. The continuity of the item in the lexicon, how it is stored in and retrieved from the "black box" are all processes that I assume. I do not see how any useful discussion of the identity of words can ignore the fact (or, less stridently, can fail to postulate) that the black box stores, and thus provides continuity to, the existence of a word. Perhaps the black box can so twist and distort a word that it loses its identity and is no longer the word it was. This might happen through the mixing of two or more different words, or it might happen in other ways. (This is a separate issue from whether the physiological apparatus of speech and hearing allows one to perform the word in a way that is locally recognizable.) But I do not see how the black box could just conjure a word out of thin air. So the links arise when I perceive the performance of a word and either take it that I recognize the word and so associate it with an item already in my lexicon, or I take it that I do not recognize it and so (perhaps) add it to my lexicon as a new word. This passage of a word from you to me is a complex operation because it turns on our fallible powers of recognition, and thus is open to the familiar perceptual errors of taking distinct perceived things to be one and taking a single thing perceived on two occasions to be two (Peter's error).

---

[64] Hawthorne and Lepore, *op cit*, p. 482.

[65] *Ibid.*

[66] *Ibid.*, p. 484.

[67] I should note a case where the formal criteria still seem to survive. The requirement (and threat) not to take the Lord's name in vain (see the Third Commandment) presumably forbids uttering the curse 'God damn it'. But it does not seem to forbid the use of the spooneristic mispronunciation 'Dad Gummit'. Taboos seem to attach to phonological types, not words. Witness the relative acceptability of 'freaking'.

[68] *Ibid.*

There remain important and interesting issues concerning common currency names, for example, how a common currency name can change its referent. I have views about this. I think it is the power of *acquaintance*, manifesting itself in what Kripke calls "speaker's reference," over the power of what convention has set to be correct usage. Interaction with the individual referred to puts us in more of a "That's how *I* use [the name]" mood, and inclines us to stick with it rather than worry about historical origins. This is especially so when the named (or misnamed) one is uncomplaining (as was Madagascar). But this is not the place to pursue this matter.

Another important matter I have said only a little about is Paderewski cases. So I will close with a final remark about word identity. If 'Londres' and 'London' are merely dialectal variants of the same name, as I suspect they are, then we have a Paderewski case before we even get to Paderewski. I have it on good authority that 'Peking' and 'Beijing' *are* the same name. I never knew that. I thought they moved the capital.

Oh my God! I am Peter.[59]

DAVID KAPLAN

University of Californa, Los Angeles

---

[59] Thanks to Sam Cumming, Eliot Michaelson, and Tamar Weber for reassuring me that these inscriptions resemble the utterances of April 2010.

MEMO TO AUTHORS

The Journal of Philosophy is published monthly, in issues that (ordinarily) contain only 56 pages of copy. This limits the length of papers normally published to no more than 7,500 words, or 18 *Journal* pages. A more precise way of determining length is by characters (including spaces): 2,850 characters per *Journal* page, or 51,300 characters per *Journal* article of 18 pages.

While the *Journal* does not publish manuscripts that mainly comment on articles published elsewhere, we do publish short responses to *Journal* articles. Submissions to this "Comments and Criticism" section should be no more than 5,000 words, or 12 *Journal* pages. *The Journal of Philosophy* does not publish papers that are primarily historical or expository, nor does it accept unsolicited book reviews.

We now ask that all submissions be made by sending an email to submissions@journalofphilosophy.org with the manuscript file attached in PDF or Word format. Please include your name, contact information, and the title of your paper in your message. All submissions will be acknowledged by email, and all subsequent correspondence regarding the manuscript will be sent via email rather than post.

Further, manuscripts must be prepared for double-blind refereeing, with any and all revealing references to the author removed, including personal acknowledgments. Any submission that does not adequately conceal the identity of its author will not be read.

Please also observe the following guidelines when preparing submissions:

(1) All material, including quotations and notes, must be typed and double-spaced, with ample margins left and right, top and bottom.
(2) No bibliographies will be published; bibliographical material should be put into notes
(3) Notes should be complete, according to *Journal* style (see any issue) Journal citations should include: the first and last name of author(s), name of journal in full, volume number, issue number, month or season, year, the article's full range of pages, and, if applicable, the specified page or pages. Book references should include: the first and last name of author, publisher, U.S. city of publication, and date Notes should be gathered at the end of the article.
(4) In quotations, spelling, capitalization, and formatting should adhere to the original.
(5) Diagrams should be integrated into the manuscript. Special symbols should be widely spaced and clearly marked.
(6) No underlining; please limit use of italics to the first appearance of technical terms and to variables
(7) The use of contractions should be limited to questions.

The editors are willing to read and evaluate a manuscript that does not meet these seven guidelines, provided the author agrees to meet them before submitting a final manuscript. Please note that the editors may ask for revision of excessively long submissions and that the final version of accepted manuscripts must be formatted as a Word file.

NEW BOOKS: ANTHOLOGIES

PALMER, R. BARTON and STEVEN M. SANDERS, eds.: *The Philosophy of Steven Soderbergh.* Lexington: Kentucky UP, 2011. ix, 318 p. Cloth: $35.00.

PIGDEN, CHARLES R., ed.: *Hume on* Is *and* Ought. New York: Palgrave Macmillan, 2010. xiv, 352 p. Cloth: $95.00.

PRIEST, GRAHAM and DAMON YOUNG, eds.: *Martial Arts and Philosophy: Beating and Nothingness.* Chicago: Open Court, 2010. xviii, 238 p. Paper: $21.95.

REATH, ANDREWS and JENS TIMMERMANN, eds.: *Kant's* Critique of Practical Reason*: A Critical Guide.* New York: Cambridge, 2010. xi, 226 p. Cloth: $90.00.

RICHARDS, TED, ed.: *Soccer and Philosophy: Beautiful Thoughts on the Beautiful Game.* Chicago: Open Court, 2010. viii, 408 p. Cloth: $21.95.

SAUNDERS, SIMON et al., eds.: *Many Worlds? Everett, Quantum Theory, and Reality.* New York: Oxford, 2010. xvi, 618 p. Cloth: $99.00.

SAVONIUS-WROTH, S.-J., PAUL SCHUURMAN, and JONATHAN WALMSLEY, eds.: *The Continuum Companion to Locke.* Baltimore: Continuum, 2010. xx, 330 p. Cloth: $180.00.

SCHMID, STEPHEN E., ed.: *Climbing: Because It's There.* Malden, MA: Wiley-Blackwell, 2010. xii, 242 p. Paper: $19.95.

SHAFER-LANDAU, RUSS, ed.: *Oxford Studies in Metaethics, Volume 5.* New York: Oxford, 2010. vi, 324 p. Cloth: $99.00, Paper: $40.00.

SHOOK, JOHN R. and PAUL KURTZ, eds.: *The Future of Naturalism.* Amherst, NY: Humanity, 2010. 272 p. Cloth: $28.98.

SORGNER, STEFAN LORENZ and OLIVER FÜRBETH, eds.: *Music in German Philosophy.* Chicago: University Press, 2011. Paper: $35.00.

STEIFF, JOSEF and ADAM BARKMAN, eds.: *Manga and Philosophy: Fullmetal Metaphysician.* Chicago: Open Court, 2010. 372 p. Paper: $19.95.

STEWART, JOHN, OLIVIER GAPENNE, and EZEQUIEL A. DI PAOLO, eds.: *Enaction: Toward a New Paradigm for Cognitive Science.* Cambridge: MIT, 2011. xvii, 481 p. Cloth: $40.00.

STUHR, JOHN J., ed.: *100 Years of Pragmatism: William James's Revolutionary Philosophy.* Bloomington: Indiana UP, 2010. ix, 215 p. Cloth: $65.00, Paper: $24.95.

TALIAFERRO, CHARLES, PAUL DRAPER, and PHILIP L. QUINN, eds.: *A Companion to Philosophy of Religion, Second Edition.* Malden, MA: Wiley-Blackwell, 2010. xvii, 760 p. Cloth: $204.95.

TALIAFERRO, CHARLES and ELSA J. MARTY, eds.: *A Dictionary of Philosophy of Religion.* Baltimore: Continuum, 2010. xxx, 286 p. Cloth: $120.00, Paper: $34.95.

TALIAFERRO, CHARLES and CHAD MEISTER, eds.: *The Cambridge Companion to Christian Philosophical Theology.* New York: Cambridge, 2010. xiv, 264 p. Cloth: $92.00, Paper: $27.99.

TENENBAUM, SERGIO, ed.: *Desire, Practical Reason, and the Good.* New York: Oxford, 2010. 251 p. Cloth: $65.00.

TIMMERMANN, JENS, ed.: *Kant's* Groundwork of the Metaphysics of Morals*: A Critical Guide.* New York: Cambridge, 2010. x, 234 p. Cloth: $85.00.

VAN TUINEN, SJOERD and NIAMH MCDONNELL, eds.: *Deleuze and the Fold: A Critical Reader.* New York: Palgrave Macmillan, 2010. xii, 263 p. Cloth: $90.00.

VANHOOZER, KEVIN J., ed.: *Remythologizing Theology: Divine Action, Passion, and Authorship.* New York: Cambridge, 2010. xix, 539 p. Cloth: $131.00.

VENTURINHA, NUNO, ed.: *Wittgenstein After His* Nachlass. New York: Palgrave Macmillan, 2010. xvii, 203 p. Cloth: $85.00.

VERNON, RICHARD, ed.: *Locke on Toleration*. New York: Cambridge, 2010. xl, 179 p. Cloth: $90.00, Paper: $19.99.

VERSTEGEN, IAN F., ed.: *Maurice Mandelbaum and American Critical Realism*. New York: Routledge, 2010. vi, 173 p. Cloth: $135.00.

WALLER, S., ed.: *Serial Killers: Being and Killing*. Malden, MA: Wiley-Blackwell, 2010. xii, 231 p. Paper: $19.95.

WARNER, MICHAEL, JONATHAN VANANTWERPEN, and CRAIG CALHOUN, eds.: *Varieties of Secularism in a Secular Age*. Cambridge: Harvard, 2010. vii, 337 p. Cloth: $45.00.

WELCH, JOHN R., ed.: *Other Voices: Readings in Spanish Philosophy*. Notre Dame: University Press, 2010. xii, 363 p. Paper: $38.00.

WHITING, DANIEL, ed.: *The Later Wittgenstein on Language*. New York: Palgrave Macmillan, 2010. xi, 243 p. Cloth: $85.00.

WITTKOWER, D. E., ed.: *Facebook and Philosophy: What's on Your Mind?* Chicago: Open Court, 2010. ix, 285 p. Paper: $19.95; *Mr. Monk and Philosophy: The Curious Case of the Defective Detective*. Chicago: Open Court, 2010. xvii, 266 p. Paper: $19.95.

ZEPKE, STEPHEN and SIMON O'SULLIVAN, eds.: *Deleuze and Contemporary Art: Deleuze Connections*. New York: Edinburgh, 2010. xi, 324 p. Cloth: $115.00, Paper: $35.00.

VIRGINIA

Fatalism in American Film Noir
Some Cinematic Philosophy
Robert B. Pippin

$24.95 | CLOTH | PAGE-BARBOUR LECTURES

"Robert Pippin examines popular movies from a philosophical perspective and does not treat them merely as an illustration of ideas but as a way of putting the ideas to the test of concrete cinematic experience. He looks at them thoughtfully and sensitively both as an inquisitive philosopher and as an attentive film critic. An original and illuminating contribution to both philosophy and film studies."
—*Gilberto Perez, Sarah Lawrence College*

ALSO OF INTEREST

The Reason of the Gift
Jean-Luc Marion
Translated by Stephen E. Lewis

$29.50 | CLOTH | RICHARD LECTURES

VIRGINIA

WWW.UPRESS.VIRGINIA.EDU

Young Epistemologist Prize - 2013

The **Rutgers Epistemology Conference** announces the eighth bi-annual prize for the best essay submitted to the Conference by a person with a Ph.D. obtained by the time of submission but not earlier than ten (10) years prior to the date of the conference. Thus, for the Conference on May 10th/11th, 2013, the Ph.D. must have been awarded prior to November 12, 2012, but not after May, 2003. The author of the prize winning essay will present it at the Conference and it will be published (along with the invited papers) in Philosophy and Phenomenological Research. The winner of the prize will receive an award of $1,000 plus all travel and lodging expenses connected with attending the Conference.

The essay may be in any area of epistemology. It must be limited to 6,000 words (including the footnotes but not the bibliography). Please send **two** copies of the paper as email attachments in a .pdf or .doc format to: YEP@philosophy.rutgers.edu

One copy must mask the author's identity so that it can be judged blindly. The second copy must be in a form suitable for publication. The email should have the subject: "YEP Submission." The email should be sent by 8 pm. (EST) on November 12, 2012. The winner of the prize will be announced by February 15, 2013.

By submitting the essay, the author agrees not to submit it to another publication venue prior to February 15, 2013 and agrees i) to present the paper at the Rutgers Epistemology Conference, ii) to have it posted on the Conference Webpage, and iii) to have it published in Philosophy and Phenomenological Research.

All questions about the Young Epistemologist Prize should be sent to the YEP email address (given above).

For more information about the Rutgers Epistemology Conference, please go to the Conference Website:

http://epistemology.rutgers.edu

CAMBRIDGE

NEW *and* NOTEWORTHY TITLES *in* PHILOSOPHY

Should Race Matter?
Unusual Answers to the Usual Questions
DAVID BOONIN
$99.00 Hb 978-0-521-76086-7
$34.99 Pb 978-0-521-14980-8
440 pp

NOW IN PAPERBACK!
Leibniz
An Intellectual Biography
MARIA ROSA ANTOGNAZZA
$29.99 Pb 978-1-107-62761-1,
664 pp

Ethics and the Media
An Introduction
STEPHEN J. A. WARD
Cambridge Applied Ethics
$90.00 Hb 978-0-521-88964-3
$28.99 Pb 978-0-521-71816-5
29[illegible] pp

Evolution and Belief
Confessions of a Religious Paleontologist
ROBERT J. ASHER
$24.99 Hb 978-0-521-19383-2
300 pp

Agro-Technology
A Philosophical Introduction
R. PAUL THOMPSON
Cambridge Introductions to Philosophy and Biology
$90.00 Hb 978-0-521-11797-5
$28.99 Pb 978-0-521-13375-3
258 pp

Prices subject to change

CAMBRIDGE CRITICAL GUIDES

Aristotle's *Nicomachean Ethics*
A Critical Guide
JON MILLER
$85.00 Hb 978-0-521-51448-4
300 pp

Nietzsche's *On the Genealogy of Morality*
A Critical Guide
SIMON MAY
$99.00 Hb 978-0-521-51880-2
356 pp

Critical Pragmatics
An Inquiry into Reference and Communication
KEPA KORTA *and* JOHN PERRY
$95.00 Hb 978-0-521-76497-1
$29.99 Pb 978-0-521-74867-4
192 pp

Political Philosophy in the Twentieth Century
Authors and Arguments
CATHERINE H. ZUCKERT
$95.00 Hb 978-1-107-00622-5
$29.99 Pb 978-0-521-18506-6
290 pp

Kant, Religion, and Politics
JAMES DICENSO
$95.00 Hb 978-1-107-00933-2
302 pp

www.cambridge.org/us/philosophy

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

from CAMBRIDGE UNIVERSITY PRESS

Philosophy and Religion
Volume 68
ANTHONY O'HEAR
Royal Institute of Philosophy Supplements
$32 99 Pb 978-1-107-61598-4
320 pp

NOW IN PAPERBACK!

Plato's Cosmology and its Ethical Dimensions
GABRIELA ROXANA CARONE
$32 99 Pb 978-1-107-65704-5
332 pp

Plato's *Charmides*
Positive Elenchus in a 'Socratic' Dialogue
THOMAS M TUOZZO
$90 00 Hb 978-0-521-19040-4
368 pp

Spinoza's Geometry of Power
VALTTERI VILJANEN
$90 00 Hb 978-1-107-00780-2
208 pp

The Cambridge Companion to Nozick's *Anarchy, State, and Utopia*
RALF M BADER *and* JOHN MEADOWCROFT
Cambridge Companions to Philosophy
$90.00 Hb 978-0-521-19776-2
$29.99 Pb 978-0-521-12002-9
332 pp

NOW IN PAPERBACK!

2 Volume Boxed Set

The Cambridge History of Eighteenth-Century Philosophy
KNUD HAAKONSSEN
$90 00 2 Pb books
978-0-521-28088-4 1,424 pp

Human Rights Law and the Marginalized Other
WILLIAM PAUL SIMMONS
$90 00 Hb 978-1-107-01007-9
250 pp

Insurrection and Intervention
The Two Faces of Sovereignty
NED DOBOS
$85 00 Hb 978-0-521-76113-0
244 pp

Liberalism and Capitalism
Volume 28
ELLEN FRANKEL PAUL, FRED D MILLER, JR, *and* JEFFREY PAUL
Social Philosophy and Policy
$36 99 Pb 978-1-107-64026-9
312 pp

Potamo of Alexandria and the Emergence of Eclecticism in Late Hellenistic Philosophy
MYRTO HATZIMICHALI
$95 00 Hb 978-0-521-19728-1
208 pp

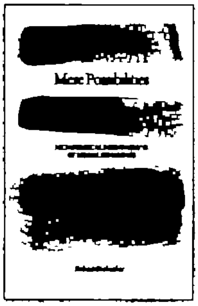
Investigating Pristine Inner Experience
Moments of Truth
RUSSELL T. HURLBURT
$99 00 Hb 978-1-107-00994-3
$36 99 Pb 978-0-521-27912-3
472 pp

Moral Status and Human Life
The Case for Children's Superiority
JAMES G DWYER
$90 00 Hb 978-0-521-76691-3
220 pp

Prices subject to change

www.cambridge.org/us/philosophy

CAMBRIDGE
UNIVERSITY PRESS



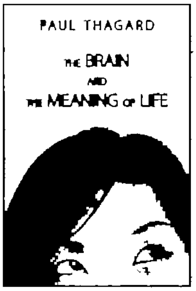
Mere Possibilities

Metaphysical Foundations of Modal Semantics

Robert Stalnaker

"This is a subtle, nuanced exploration of the metaphysical basis for talk of possible worlds and its far from straightforward relation to the version of formal semantics known as 'possible worlds semantics,' both of which have been immensely influential in recent decades. Robert Stalnaker has been one of the major players in debates on these matters, and this book contains significant further developments of his ideas."
—Timothy Williamson, University of Oxford

Cloth $45.00 978-0-691-14712-3



The Brain and the Meaning of Life

Paul Thagard

"[Thagard] offers a tightly reasoned, often humorous, and original contribution to the emerging practice of applying science to areas heretofore the province of philosophers, theologians, ethicists, and politicians: . . . What is the source of the sense of self? What is love? What is the difference between right and wrong, and how can we know it? What is the most legitimate form of government? . . . Thagard employs the latest tools and findings of science in his attempts to answer these (and additional) questions."
—Michael Shermer, *Science*

Paper $19.95 978-0-691-15440-4

Meaning in Life and Why It Matters

Susan Wolf

"Susan Wolf is one of the clearest, most thoughtful, and most incisively elegant writers in contemporary ethics. She has an uncanny knack for putting her finger on important points and expressing them in ways that capture the imagination. In this book, she develops her ideas about meaningfulness in life with considerable subtlety, creating a work of genuine depth and importance."
—John Martin Fischer, University of California, Riverside

Paper $19.95 978-0-691-15450-3

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 10
OCTOBER 2011

page

535 *A Distinction within Egalitarianism* Alan Carter
555 *Concepts and Abilities in Anti-Individualism* Endre Begby

COMMENTS AND CRITICISM

576 *An Insubstantial Externalism* Axel Arturo Barceló Aspeitia

BOOK REVIEWS

583 G. A. Cohen: *Rescuing Justice and Equality* Japa Pallikkathayil

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Teddy Seidenfeld, Gila Sher, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

---

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2011 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 10, OCTOBER 2011

A DISTINCTION WITHIN EGALITARIANISM

It is often taken for granted that egalitarians value a completely equal distribution at any moment in time of whatever it is that they believe should be distributed equally. But, I argue, an alternative form of egalitarianism may be contrasted with the variety to which it is often presupposed egalitarians subscribe. Call this alternative form "teleological egalitarianism." Roughly, teleological egalitarianism evaluates momentary distributions according to their contribution to an overall distribution across complete lives. In contradistinction to teleological egalitarianism, call what has been taken to be the traditional form "nonteleological egalitarianism." I shall clarify the distinction between teleological and nonteleological egalitarianism in due course.

Before I do, however, it is worth noting that neither teleological egalitarianism nor, in my view, and contrary to widespread assumptions, the classical variety is committed to the pursuit of complete equality. Rather, I contend, egalitarianism is opposed to certain substantive inequalities.[1] Nevertheless, in order to gauge whether or not any inequality is substantive, one requires some conception of an equal distribution to serve as one's yardstick. For it is only by reference to such a distribution that the degree of inequality can be assessed. Armed with an appropriate conception of equality, one can then assess the divergence of some distribution from an equal one.

Interestingly, there are cases where an equal distribution that would be viewed by a nonteleological egalitarian as possessing value would be viewed by a certain kind of teleological egalitarian as possessing disvalue. Likewise, there are cases where an unequal distribution that would be viewed by a nonteleological egalitarian as possessing disvalue

---

[1] See Alan Carter, "Analytical Anarchism: Some Conceptual Foundations," *Political Theory*, xxviii, 2 (April 2000): 230–53, especially p. 231, where I follow John Baker, *Arguing for Equality* (New York: Verso, 1987), pp. 4–5.

0022-362X/11/0810/535–554

would be viewed by a certain kind of teleological egalitarian as possessing value. The distinction between teleological and nonteleological egalitarianism is, therefore, not a trivial one.

I

Even though I hold that egalitarianism is best construed as the opposition to certain substantive inequalities, in what follows I shall, for ease of exposition, often argue as if egalitarians seek to bring about a completely equal distribution. Before proceeding further, it might be fruitful to allay one possible fear. Egalitarians presumably "see" value in equal distributions or, perhaps more accurately, "see" disvalue in unequal ones. But do they have good reasons for so doing? For example, it might be argued that, given the prevalent values of western liberal democracies, egalitarianism, unlike Rawlsian political philosophy, is unmotivated.[2] But Rawls's original argument for his two principles of justice appears to take equality as the default position.[3] And it would be odd to regard inequalities as requiring justification if equality were of no moral value. Indeed, it would seem likely that equality would have to be of at least some intrinsic value if inequalities must be provided with a justification for them to be viewed as morally permissible.

Moreover, it would seem that some notion of equality lies at the basis of our common moral views. For if two people were to walk into my room and I were simply to attack one verbally while praising the other, my behavior would no doubt be deemed immoral. But if it were known that one of the two had pushed a child into the path of an oncoming car and the other, at the risk of her own life, had leapt in front of the car to save the child, then my otherwise discriminatory behavior would no longer be presumed to be immoral. This suggests that we hold as a basic moral principle that persons ought to be treated equally unless there is a morally significant[4] and morally relevant difference[5] between them that justifies unequal treatment.[6] And this suggests that unjustified inequalities are of at least some intrinsic disvalue, which further seems to suggest that equality possesses some intrinsic value. All that an egalitarian need do to motivate her egalitarianism is argue that there are no morally significant differences between persons—differences which are also morally relevant to the inequality in question—that justify the inequalities she denounces. If she can succeed in so arguing, then it seems that she will have provided grounds for holding that the inequality she opposes is, at least to some degree, of intrinsic disvalue.[7]

---

[2] Samuel Scheffler, "What Is Egalitarianism?" *Philosophy and Public Affairs*, XXXI, 1 (Winter 2003): 5–39.

[3] See John Rawls, "Justice as Fairness," in Peter Laslett and W. G. Runciman, eds., *Philosophy, Politics and Society: Second Series* (Oxford: Basil Blackwell, 1962). "Justice as Fairness" was first published in *The Philosophical Review*, LXVII, 2 (1958): 164–94.

[4] See, for example, Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Indianapolis: Hackett, 1981), pp. 209, 380.

[5] For one argument that seeks to establish that differences should be relevant to the inequalities they are employed to justify, see Bernard Williams, "The Idea of Equality," in Laslett and Runciman, eds., *op. cit.*

[6] Indeed, without some such presumption, it is difficult to make much sense of the near universal agreement between metaethicists that moral properties supervene upon nonmoral ones.

On the other hand, egalitarians might denounce inequalities for purely instrumental reasons.[8] In what follows, I am not concerned with whether an egalitarian sees intrinsic or instrumental value in equality. Nor, conversely, am I concerned with whether she sees intrinsic or instrumental disvalue in inequality. I shall simply assume that the egalitarian sees some value in equality or some disvalue in inequality, regardless of whether it is intrinsic or merely instrumental. And I shall refer to the value that she sees in a distribution as "distribution value."

II

Of course, egalitarians do not always see eye to eye on which inequalities ought to be ameliorated. Indeed, egalitarian philosophers have famously disagreed over what should be distributed equally or, better in my view, what should be distributed less unequally. For example, while one might presume that egalitarians are preoccupied with the distribution of welfare, some egalitarians have, rather, been concerned with the distribution of resources,[9] basic capabilities,[10] opportunities for welfare,[11] access to advantage,[12] or some combination of the above.[13] My present aim is not to adjudicate this dispute. It is to distinguish between two ways an egalitarian might value some *momentary* distribution with respect to an equality, or disvalue it with respect to an inequality, regardless of what the egalitarian thinks ought to be distributed in a certain way. Hence, while clarifying the distinction that I wish to focus upon, I do not want to prejudge the discussion concerning what, if anything, ought to be distributed less unequally.

---

[7] Here I have, in part, attempted to motivate the value of equality. Elsewhere, I have attempted to motivate the value of others' freedom. See Carter, "Morality and Freedom," *Philosophical Quarterly*, LIII, 211 (April 2003): 161–80. I have also attempted to motivate the value of others' well-being in Carter, "Humean Nature," *Environmental Values*, IX, 1 (2000): 3–37.

[8] This is a consideration that I return to in section IV, below

[9] The most famous exponent of equality of resources is Ronald Dworkin. See his "What Is Equality? Part 1: Equality of Welfare," *Philosophy and Public Affairs*, X, 3 (Summer 1981): 185–246; and especially his "What Is Equality? Part 2: Equality of Resources," *Philosophy and Public Affairs*, X, 4 (Fall 1981) 283–345.

[10] Amartya K. Sen is the pioneering exponent of "basic capabilities equality." See his "Equality of What?" in *Choice, Welfare and Measurement* (Cambridge: MIT, 1982)

[11] This is the egalitarian metric preferred by Richard J Arneson in his "Equality and Equal Opportunity for Welfare," *Philosophical Studies*, LVI, 1 (May 1989): 77–93.

[12] Jerry Cohen advocates this particular egalitarian metric See G. A. Cohen, "On the Currency of Egalitarian Justice," *Ethics*, XCIX (July 1989): 906–44.

[13] I argue for a pluralist approach in Carter, "Value-Pluralist Egalitarianism," this JOURNAL, XCIX, 11 (November 2002): 577–99

To this end, let us employ "*G*" to refer to whatever the egalitarian in question believes ought to be distributed as equally as possible, practicable, or desirable. Thus, *G* might be welfare, resources, basic capabilities, opportunities for welfare, access to advantage, some combination of these,[14] or some metric yet to be identified. But there is a further way in which egalitarians might differ—one that cuts across all of the differences with respect to what they hold ought to be distributed less unequally. For egalitarians might be concerned with a synchronic distribution, as in "the simultaneous segments view," or instead they might be concerned with some overall outcome, as in "the complete lives view."[15] And if an egalitarian subscribes to the complete lives view (where lives as a whole, rather than segments of lives, are to be compared), then she may evaluate a momentary distribution teleologically. In this article, my principal focus of concern is the distinction between teleological and nonteleological egalitarian evaluations.

Utilitarianism, of course, constitutes a clear case of a teleological theory. Utilitarians are concerned with how an action, say, actually contributes or will probably contribute to some overall outcome.[16] According to their normative theory, bringing about the good (which is understood differently by different varieties of utilitarianism), or a high probability of bringing about enough of it, is what makes an action right. Thus, failing to bring about this end when an alternative action would have done so, or choosing an action with a lower expected utility than some other action, is what makes an action wrong. Now, it is usually presumed that egalitarians are nonteleological in the way they evaluate any distribution. In contrast to utilitarianism,

---

[14] For an overview of the arguments in favor of each of these metrics, see Carter, "The Quest for an Egalitarian Metric," *Critical Review of International Social and Political Philosophy*, VII, 1 (Spring 2004): 94–113.

[15] On the distinction between the simultaneous segments view and the complete lives view, see Dennis McKerlie, "Equality and Time," *Ethics*, XCIX, 3 (1989): 475–91.

[16] For the distinction between "consequentialism based on actual results" and "consequentialism based on probable results," see Tim Mulgan, *The Demands of Consequentialism* (New York: Oxford, 2001), p 33. In what follows, I adopt Roger Crisp's terminology, and refer to the former variety of consequentialism as "actualist" and the latter variety as "probabilist." See Crisp, *Mill on Utilitarianism* (New York: Routledge, 1997), pp. 99–100. I shall deploy these terms to distinguish between actualist and probabilist forms of teleological egalitarianism.

egalitarianism is clearly nonteleological when it simply holds that an equal distribution at some particular moment in time of whatever egalitarians are presumed to want to equalize (*G*) possesses intrinsic distribution value. Consequently, on such a view, if at any moment in time a significantly unequal distribution of *G* obtains, then that distribution manifests disvalue. If an egalitarian subscribes to the simultaneous segments view, rather than to the complete lives view, then she will certainly be a nonteleological egalitarian.

But the simultaneous segments view appears to lead to some unwelcome implications. For imagine that resources are what the egalitarian holds should be distributed equally. Someone might prefer to save her share of resources and consume them later. But when the resources she has saved are combined with her later share, she will then possess more resources than those persons who chose not to save. An unequal distribution of resources will thus obtain. Egalitarianism might appear unattractive if it were to object, on the grounds that an inequality would result, to a person preferring to enjoy her resources later rather than at the same time as everyone else. Equally, welfare might be what the egalitarian holds should be distributed in an equal fashion. But someone easily might prefer to enjoy less welfare now and more later. When that person so acts that some of her welfare is postponed, her action may thereby result in her enjoyment of more welfare at some future moment than is enjoyed by those persons who did not defer their gratification. An unequal distribution of welfare will then obtain. Egalitarianism might appear unattractive if, on the grounds that an inequality would result, it were to object to a person enjoying more welfare later rather than at the same time as the others, when that later enjoyment was voluntarily bought at the cost of an earlier deferment of gratification.

In order to avoid this seemingly counterintuitive implication, an egalitarian might argue that it is the total amount of *G* that one enjoys over the course of one's life that matters. In other words, there are reasons for accepting, in contrast to the simultaneous segments view, the complete lives view. On such a view, the equality that counts might concern the total amount of *G* that each person ends up enjoying. Hence, it is irrelevant, on the complete lives view, whether someone enjoyed more *G* than anyone else earlier in her life, as long as she enjoyed that amount less than everyone else later on. What is required, then, is that, by the end of his or her life, each person will have enjoyed the same total quantity of *G* as everyone else.

But if it were known that someone was going to live a shorter life than others, it might seem to follow that she should, on average, receive more at any time than anyone else. For example, if person

A were to live 40 years and person *B* were to live 80 years, then, on average, it might be thought that *A* should enjoy twice as much *G* as *B*. But this is likely to offend egalitarian intuitions. For one person to enjoy twice as much at every moment of her life as another person might seem highly inegalitarian.[17]

This possibly counterintuitive implication could be avoided by focusing upon the average level of *G* that individuals enjoy at any moment in time. Let G^{ao} equal the total quantity of *G* that a person *x* enjoys over the whole of her or his life divided by the length of time that she or he lives. If G^{ao} is the same for each person, then what we might call "an equal average overall outcome" will have obtained. If an egalitarian were to value such an equal average overall outcome, rather than an equal distribution at any moment in time, then she could avoid both the seemingly counterintuitive problem of the short-lived having to be given considerably more of *G* at any moment in time than the longer-lived, and the seemingly counterintuitive problem of having to prevent some from deferring their enjoyment of whatever it is that she holds ought to be equalized.

But how might such an egalitarian ascertain the value of a distribution at any moment in time? If she does not want to condemn an inequality that obtained because some individual deferred his enjoyment of *G*, she could, say, evaluate a temporary distribution in terms of its contribution to the average overall outcome. If it contributed to an equal average overall outcome, then the temporary distribution would possess value in proportion to its contribution, and if it contributed instead to an unequal average overall outcome, then the temporary distribution would posses disvalue in proportion to its contribution.

Any variety of egalitarianism that evaluates a temporary distribution in terms of its contribution, or expected contribution, to some specific overall distributional outcome, I refer to as a form of teleological egalitarianism. Nonteleological varieties of egalitarianism, by contrast, evaluate a distribution simply in terms of whether, and to what extent, it is unequal. If a momentary distribution is equal, then it possesses value. If it is unequal, it possesses less value. And if it is highly unequal, it possesses considerable disvalue. By contrast, according to the teleological egalitarian who is concerned solely with actual results—the actualist teleological egalitarian—an unequal momentary distribution would posses value if, because of an unequal distribution at some other time, it contributed to an equal overall outcome. That equal overall outcome might be construed in terms of the average level of *G* enjoyed, the total amount enjoyed, or some compromise between both the total amount and the average level of *G* enjoyed.[18]

---

[17] It might be thought unfair that some should live only half as long as others. One approach might be to regard unequal lifespans (which might be viewed as demanding compensation for the short-lived) and unequal distributions of *G* as both possessing disvalue, and then engage in trade-offs in order to generate the optimal outcome. Pursuing this line of reasoning would take us too far from our present concern. However, see Carter, "Towards a Multidimensional, Environmentalist Ethic," *Environmental Values*, XX, 3 (2011): 347–74 and Carter, "Some Groundwork for a Multidimensional Axiology," *Philosophical Studies*, CLIV (2011): 389–408 for discussions of how value trade-offs might fruitfully be modeled.

In order to make clearer this notion of teleological egalitarianism, imagine a world limited to two people with the same lifespan. In the imagined world, even though at any particular time they enjoy different quantities of *G*, by the end of their lives they will have enjoyed the same quantity of it. The complete lives egalitarian would be concerned only with such an overall distribution. If the egalitarian were to judge the temporary unequal distributions solely in terms of their contribution, or their expected contribution, to the overall equal distribution, and thus valued them positively, then she would be valuing teleologically. It is just such an evaluator that I refer to as a "teleological egalitarian." She is to be contrasted primarily with an egalitarian who is unconcerned with the overall distribution over time, and who is, instead, concerned solely with the inequalities that obtain at any particular moment. If the latter kind of egalitarian views those inequalities as of negative value in themselves, regardless of their contribution to an equal final outcome with respect to the complete lives of the two persons residing in that world, then she would be valuing nonteleologically. Such an evaluator I refer to as a "nonteleological egalitarian."

III

Nonteleological egalitarianism is quite familiar to us. We have all heard criticisms of some momentary inequality that nevertheless has contributed to an equality overall. In the next section, I consider an epistemological reason that an egalitarian might provide for being concerned with such an inequality. But teleological egalitarianism is not a widely recognized view. Hence, further clarification of how the teleological egalitarian evaluates might prove fruitful. To this end, consider a more familiar case of valuing teleologically. Consider how an actualist, as opposed to a probabilist,[19] utilitarian would evaluate an action's rightness or wrongness in the following situation. Imagine that a local population's food supply has become contaminated. Were the food to be eaten, the whole population would become sick, though not fatally. Further imagine that we can only replace that food if we first destroy the existing stocks (perhaps to safeguard the replacement food from becoming contaminated, and if we were not to take that precaution, the replacement food, in scarce supply, would simply be wasted). Finally, imagine that the following three courses of action are available: First, we could distribute the contaminated food. No one would die, but the whole local population would suffer from a very unpleasant illness. Second, we could destroy the contaminated stocks without replacing them. The result would be that the whole local population would starve to death. Third, we could first destroy the existing food stocks and then replace them with uncontaminated food. The whole population would thus be well fed, and no one would become ill.

---

[18] McKerlie suggests the possible desirability of a compromise between a total and an average egalitarian view. See McKerlie, *op. cit.*, p. 477.

[19] See note 16

Given these alternatives, how might an actualist utilitarian answer the question: was it right or wrong to have destroyed the food stocks? Because she is concerned with the overall outcome, she will be unable to answer that question definitively unless further information is forthcoming: namely, was replacement, uncontaminated food later provided? If it was, then a utilitarian of this sort will judge that it was right to have destroyed the stock of contaminated food. If replacement food was not provided, then she will judge that it was wrong to have destroyed the food stocks. In other words, in her view, the rightness or wrongness of destroying the contaminated food cannot be ascertained without reference to the final, overall outcome. For the actualist utilitarian, it would have been right not to have destroyed the food stocks if the replacement food were not provided in good time, but wrong not to have destroyed them if the replacement food were later supplied conditional upon the contaminated foods stocks having been destroyed. This is because the acquisition of the replacement food would have resulted in a better overall outcome than retaining the contaminated food, and retaining the contaminated food would have resulted in a better overall outcome than the local population being left with no food at all.

In a nutshell, if one values purely teleologically with reference to the actual results, then the value of a momentary event, say an action, is dependent upon the contribution of that event to the final, overall outcome. For a teleological evaluator of this sort, the value of an event cannot be ascertained independently of the overall outcome. In terms of our example, for anyone who values teleologically and with reference to the actual results, the rightness or wrongness of destroying the food stocks cannot be ascertained in abstraction from the later provision of or failure to provide uncontaminated food.[20] Hence, the value of a momentary event, say an action, is dependent upon the later or earlier events that also contribute to the final, overall outcome.

---

[20] Note that the ultimate appraisal of the rightness or wrongness of the action is not identical to judging what one ought to do. In deciding in advance whether or not to

A teleological egalitarian values similarly to the utilitarian, except that she is concerned with the final, overall distribution of *G* across time, rather than with the total quantity of well-being, preference satisfaction, or whatever the utilitarian holds should be maximized. The value of a temporary distribution, in the view of an actualist teleological egalitarian, is therefore dependent upon the contribution of that distribution to the final, overall distribution over time. Its value cannot be ascertained independently of that overall distribution. Hence, to such an egalitarian, the value of a temporary distribution is dependent upon the later or earlier distributions that also contribute to the final, overall distribution. In this way, for the actualist teleological egalitarian, a temporary inequality might have positive value. It would have such value if a later, reverse inequality led to a final, overall equality. Similarly, a temporary equality might have disvalue. This would be the case if a later, inevitable inequality led to a final, overall inequality, whereas a temporary inequality (as opposed to the temporary equality that obtained) instead would have contributed to a final, overall equality.

The nonteleological egalitarian, by contrast, does not judge the value of temporary distributions in terms of the final, overall outcome towards which they contribute. Rather, she judges them directly. Temporary distributions possess value if they instantiate an equality, and they possess disvalue to the extent that they instantiate a substantive inequality. If we must choose between nonteleological and teleological egalitarianism, which should we prefer?

IV

Well, what grounds might we have for preferring nonteleological over teleological egalitarianism? An egalitarian might have a number of grounds for valuing an equal distribution of *G* at all times. For example, she could find instrumental disvalue in temporary inequalities

---

destroy the contaminated food stocks, all utilitarians would factor in the probability of the uncontaminated food being provided. If there were an extremely high probability of the replacement food being distributed, then a utilitarian might well conclude that the contaminated food stocks ought to be destroyed. But for an actualist utilitarian, if it later turned out that the replacement food was not forthcoming, she would then judge that if anyone had destroyed the contaminated food stocks, he or she would, in fact, have acted wrongly. Consequently, I am in disagreement with McKerlie when he seems to claim that utilitarianism is an exclusively forward-looking doctrine. See McKerlie, *op. cit.* It might be exclusively forward-looking with respect to the *summum bonum*, but this is not necessarily so with respect to rightness and wrongness.

because they pose problems for the satisfaction of human needs, or for mutual respect, or for the maintenance of a genuine sense of community—things she might value intrinsically. Or she might value one, or some combination, of them instrumentally because it is required for what she considers to be a good life, or for what she regards as human flourishing—either of which she might well value intrinsically. More precisely, it could be argued that if one values, either intrinsically or instrumentally, the satisfaction of human needs, then one should oppose inequalities at every moment, for, the argument might proceed, the diminishing of inequalities is the most appropriate means to their satisfaction.[21] Alternatively, it could be argued that if one values, either intrinsically or instrumentally, mutual respect, then one should oppose inequalities at every moment, for inequalities usually undermine it.[22] Or it could be argued that if one values, either intrinsically or instrumentally, a sense of community, then one should oppose inequalities at every moment, for when people are unequal genuine community seems to be rendered more difficult to attain.[23] But if we take any of these as a reason for valuing an equal distribution of *G* at all times, rather than a distribution of *G* that is equal overall but not equal at any one time (regardless of whether the extension of *G* is confined to well-being), then this is likely to be the case because we are not pure egalitarians[24] but also value, say, community or human flourishing.

However, there is another reason why one might value an equal distribution of *G* at all times. If, in a two-person world, one individual were to enjoy a high level of *G* earlier and a low level later, while the other individual were to enjoy a high level of *G* later and a low level earlier, then that might seem fair enough to many, especially if the temporary inequalities were due solely to the individuals' choices. But the egalitarian is unlikely to know in advance that both

---

[21] See Baker, *op. cit.*, chapter 2.

[22] See *ibid.*, chapter 3.

[23] See *ibid.*, chapter 4.

[24] On the distinction between "pure egalitarians" and "pluralist egalitarians," see Derek Parfit, *Equality or Priority? The Lindley Lecture* (Lawrence: Kansas UP, 1995), p. 5. Parfit argues: "Suppose. .that the people in some community could all be either (1) equally well off, or (2) equally badly off. The Principle of Equality does not tell us that (2) would be worse. This principle is about the badness of inequality; and, though it would be clearly worse if everyone were equally worse off, our ground for thinking this cannot be egalitarian." *Ibid.*, p. 4. Hence, "equality cannot plausibly be our only value. If we are Egalitarians, we must hold some more complicated view." *Ibid.*, p. 22. Thus, as few, if any, egalitarians would hold that (1) is not better than (2), it would seem implausible to regard the majority of egalitarians as genuinely pure rather than pluralist ones. For some reasons to prefer value-pluralist egalitarianism over prioritarianism, see Carter, "Some Groundwork for a Multidimensional Axiology"

individuals will live long enough for the overall enjoyment of *G* to equalize. For all the egalitarian knows, both individuals might die young. If they do, the one who enjoyed a high level of *G* at a young age will have enjoyed a lot more of it than the one who, at a young age, enjoyed a low level of *G*. Given her epistemological limitations, in order to prevent such a possibility, the egalitarian might see the need to ensure that, at any moment in time, individuals, as far as is practicable, enjoy the same level of *G*. Hence, there seems to be nothing necessarily irrational in the egalitarian preferring that all synchronic distributions remain equal. Hence, it would seem that we have reason for valuing equality within simultaneous segments in contrast to valuing equal complete lives.

On the other hand, "seeing" the value of equal simultaneous segments might simply be a moral intuition held by certain egalitarians. And an egalitarian might well "see" the intrinsic disvalue of an unequal synchronic distribution because she can find no compelling morally significant and morally relevant difference that would justify the inequality in question.

V

The above notwithstanding, are there any grounds for preferring teleological over nonteleological egalitarianism? Teleological egalitarianism, as I have portrayed it, presupposes the complete lives view. If certain temporary inequalities are inevitable, then the appropriate anticipated or later compensation to any person who temporarily enjoys less *G* than others would make the distribution of *G* more equal overall with respect to complete lives. And this might be the most equal state of affairs that is practicable.

On the other hand, temporary inequalities, even if not inevitable, might be desirable in some respect. In section II, it was noted that a person might prefer to have more *G* later rather than earlier, or *vice versa*. And a pluralist egalitarian might value freedom intrinsically. Or she might value freedom instrumentally, perhaps because it is necessary for what she considers a good life or because it is needed for what she considers human flourishing—either of which she might value intrinsically. Put another way, it could be argued that if one values freedom, either intrinsically or instrumentally, as well as equality, then one should value equal complete lives because, the argument might proceed, the choice of when to enjoy most *G* is a valuable freedom, so long as one does not enjoy a greater total quantity of *G* than anyone else over the course of one's life. Thus, the complete lives view might be thought to maximize freedom within a specific egalitarian constraint. In short, preserving equality with respect to complete lives might be how best to maximize freedom while maintaining equality in some form. But if we take this to constitute a reason for preferring a distribution that is equal overall regarding complete lives, but not equal at any one time (regardless of whether the extension of *G* is confined to well-being), to an equal distribution of *G* at all times, then, of course, it will be because we are not pure egalitarians but pluralists who also value freedom.

Alternatively, we simply might think that it is not rational to prefer a higher level of well-being earlier in one's life to later, or later to earlier, or either to an equal level of well-being throughout one's life, or *vice versa*, just so long as the total quantity of well-being is the same in each case.[25] Thus, one might conclude that, in any two-person world, as long as both individuals enjoy the same overall level of well-being over the course of their lives, neither has grounds for complaint. And hence, it might be thought, an egalitarian has no rational grounds for preferring a world in which both individuals enjoy the same level of well-being at the same time over a world in which one enjoys a higher level of well-being earlier and the other enjoys it later. In other words, it might be argued that an egalitarian has no rational grounds for preferring a world in which equality obtains at each moment to one in which at no moment does equality obtain, but where the final, overall outcome is an equal one.[26] If this argument were cogent, then it would seem we have good reason for valuing complete lives that are equal overall (in contrast to valuing equality within simultaneous segments).

On the other hand, "seeing" the value of equality across complete lives might simply be a moral intuition held by certain egalitarians. And when temporary inequalities are inevitable, and when they could be compensated for later, and, thus, when equality with respect to complete lives is the only practicable form of equality, an egalitarian might well "see" the intrinsic disvalue of an unequal distribution across complete lives because she can find no compelling morally significant and morally relevant difference that would justify such a preventable, overall inequality.

VI

In the previous section, we encountered the claim that an egalitarian has no rational grounds for preferring a world in which people enjoy the same level of well-being at the same time over a world in which

---

[25] See, for example, Michael Huemer, "Non-egalitarianism," *Philosophical Studies*, CXIV, 1-2 (May 2003): 147-71.

[26] See *ibid.*, pp. 151-53.

some enjoy a higher level of well-being earlier and the rest enjoy it later, just so long as, by the end of their lives, everyone enjoys the same quantity of well-being. This is because, it might be argued, no one has grounds for complaint in such a situation. But this could be taken further. For it might be held that this entails that the simultaneous segments view ought to be rejected. However, all of this seems to presuppose that an egalitarian can justifiably "see" disvalue in an inequality only if someone suffering an inequality has grounds for complaint on the basis of her being disadvantaged relative to other people. But if there is no compelling morally significant and morally relevant difference that would justify the inequality in question, then an egalitarian still might "see" disvalue in that inequality, even if no particular person could legitimately complain on the basis of suffering a relative disadvantage. Moreover, if an egalitarian valued community, and if temporary inequalities undermined it, she might well have reason for "seeing" disvalue in those inequalities, even when no one suffering from that inequality had grounds for complaint on the basis of her being disadvantaged relative to other people, given that all such inequalities would eventually be compensated.[27] Hence, even if there were no rational, egoistic grounds for preferring an equal distribution of *G* throughout one's life to the same total quantity of *G* being distributed unequally through the course of one's life, and even if this provides reason for valuing in the manner of a teleological egalitarian, it nevertheless fails to defeat nonteleological egalitarianism.

In a nutshell, there are arguments that count in favor of teleological modes of evaluating, yet fail conclusively to defeat nonteleological modes. But there are also arguments that count in favor of nonteleological modes of evaluating, yet fail conclusively to defeat teleological modes. For example, recall the epistemological consideration deployed in section IV in favor of the mooted rightness of maintaining an equal distribution at all moments in time. This might appear, *prima facie*, to support a judgment that nonteleological egalitarianism is preferable to the teleological variety. However, it does not, in fact, establish any such thing. This is because teleological egalitarians also can deploy that epistemological consideration to justify the preservation of equality between persons at every moment in time. For maintaining an equal distribution at all moments in time may be instrumental to ensuring that an equal overall outcome is obtained, the latter being what the teleological egalitarian might intrinsically value. Hence, because both teleological and nonteleological egalitarians could have reason for valuing equal temporary distributions, it seems difficult to provide conclusive grounds for preferring nonteleological to teleological egalitarianism or *vice versa*.[28]

---

[27] Thus, I take myself to be in disagreement with Larry Temkin, whose discussion of inequality seems to presuppose that inequalities only matter morally if a person suffering from an inequality has grounds for complaint on the basis of her being disadvantaged relative to other people. See Temkin, *Inequality* (New York: Oxford, 1996).

Fortunately for egalitarians, they may not have to choose between teleological and nonteleological varieties of egalitarianism. For, a reason for valuing what actualist teleological egalitarianism focuses upon does not have to constitute a reason for preferring actualist teleological to nonteleological egalitarianism, and *vice versa*. This is because an egalitarian could value intrinsically both what nonteleological egalitarianism and what actualist teleological egalitarianism focus upon: the distribution value of equality within simultaneous segments and the distribution value of equal complete lives, respectively.

Elsewhere, I have argued that the development of an adequate normative theory often goes astray because of a widespread presumption of value monism.[29] Consider standard versions of utilitarianism. They might seem plausible as normative theories when applied to a fixed number of people, but they encounter serious difficulties when employed to ascertain the optimum human population. For example, according to Derek Parfit, classical hedonistic utilitarianism, which subscribes to the Total View (where we are enjoined to bring about the world with the greatest total quantity of happiness), implies that, "[c]ompared with the existence of very many people—say, ten billion—all of whom have a very high quality of life, there must be some much larger number of people whose existence, if other things are equal, would be *better*, even though these people would have lives that are barely worth living."[30] This is the so-called Repugnant Conclusion, and it appears to be a highly counterintuitive implication of the Total View.

The Repugnant Conclusion can, of course, be sidestepped by instead adopting the Average View (where we are enjoined to bring about the world with the highest level of average happiness).[31] But this leads immediately to the Problem of the Ecstatic Psychopath:[32] Imagine a world containing a solitary person with no need for others—say, a psychopath—and who happens to be ecstatic. While this world may have the highest level of average happiness, it seems highly counterintuitive to regard it as the best possible world.

---

[28] But what of the person who prefers to enjoy less *G* now and more later? If she thereby loses out because she dies before the anticipated increase in *G* is enjoyed, we might regard this as bad option luck, and not something with which an egalitarian need concern herself. For the distinction between "option luck" and "brute luck," see Dworkin, "What Is Equality? Part 2: Equality of Resources," p. 293. However, it does seem likely that there are limits to how bad option luck can be and yet be completely discounted. See Baker, *op. cit.*, p. 67; Elizabeth Anderson, "What Is the Point of Equality?" *Ethics*, CIX, 2 (1999): 287–337, especially pp. 295–96; and Scheffler, *op. cit.*, p. 33.

[29] See Carter, "Towards a Multidimensional, Environmentalist Ethic."

[30] Parfit, "Overpopulation and the Quality of Life," in Peter Singer, ed., *Applied Ethics* (New York: Oxford, 1986), pp. 145–64, at p. 150.

Why do we regard the world of the Repugnant Conclusion to be worse than a world containing ten billion people, all of whom enjoy a very high quality of life? Surely it can only be because of the low level of average happiness in such a world. And why do we also regard the world of the Ecstatic Psychopath to be worse than a world containing ten billion people, all of whom enjoy a very high quality of life? It might be because we value a large number of worthwhile lives, or because we value a large total quantity of happiness, or because of some other value which we hold dear that the world of the Ecstatic Psychopath fails to satisfy. But one thing is certain: it is not because of the level of average happiness within that world. Hence, anyone who regards both the Repugnant Conclusion and the Problem of the Ecstatic Psychopath as challenges to utilitarianism has good reason to regard value monism as false, for the value that the world of the Repugnant Conclusion fails to satisfy adequately is clearly a different value from the one that is inadequately satisfied within the world of the Ecstatic Psychopath. At least two different values appear to be required if we are to make sense of our intuitive responses to both of these problems.[33] Thus, sidestepping both challenges seems to require of consequentialists that they recognize more than one value.

But if utilitarians implicitly recognize more than one value when recognizing both the Repugnant Conclusion and the Problem of the Ecstatic Psychopath as threats to their theory, then why presume that egalitarians must be value monists?

---

[31] Another way of sidestepping the Repugnant Conclusion is to deny that moral evaluations are transitive. But this seems far too high a price to pay, especially if an alternative solution can be found.

[32] See Carter, "Towards a Multidimensional, Environmentalist Ethic."

[33] If there are plural values, and we have just witnessed the need to recognize more than one value, then the question which arises immediately is how to trade them off. Elsewhere, I have proposed a normative technology for identifying, at least in principle, the best practicable world when a plurality of values needs to be accommodated. See Carter, "Value-Pluralist Egalitarianism," and, in particular, "Towards a Multidimensional, Environmentalist Ethic." Such a technology could be utilized to rescue consequentialism from the impasse that the problem of ascertaining the optimum human population has created, for it enables both the Repugnant Conclusion and the Problem of the Ecstatic Psychopath to be sidestepped simultaneously. On how such a normative technology could be applied to environmental ethics, see *ibid.*

VII

Interestingly, it seems to me that many egalitarians, even if they have yet to draw an explicit distinction between teleological and nonteleological egalitarianism, adhere to both the value focused upon by nonteleological egalitarianism and the value focused upon by teleological egalitarianism concerned with actual results. To see why this might be so, consider an egalitarian who intuitively "sees" intrinsic disvalue in inequalities, and imagine a world, W_1, where everyone lives the same number of years and enjoys the same amount of G at any moment. W_1 is an equal world on both the simultaneous segments view and the complete lives view. Because it is an equal world on the complete lives view, any moment will contribute to an equal overall outcome. Hence, according to actualist teleological egalitarianism, every moment has positive distribution value. And because at every moment everyone enjoys the same quantity of G, every moment also has positive distribution value according to nonteleological egalitarianism.

Now imagine a different world, W_2, where everyone lives the same number of years and will have enjoyed the same total quantity of G by the end of their lives, but where two classes enjoy vastly different quantities of G at any single moment. Imagine that halfway through their lives, the rich class suddenly becomes the poor class, and, simultaneously, the poor class suddenly becomes the rich class. To make this more precise, imagine that, at any moment, each individual in one class enjoys 250 times as much G as any individual in the other class. W_2 is an unequal world on the simultaneous segments view, but it is an equal world on the complete lives view. Because it is an equal world on the complete lives view, any moment will contribute to an equal overall outcome. Hence, according to actualist teleological egalitarianism, every moment has positive distribution value. But because at every moment the members of one class enjoy a far greater quantity of G than the members of the other, every moment has negative distribution value according to nonteleological egalitarianism.

Finally, imagine yet another world, W_3, where everyone lives the same number of years, where individuals enjoy vastly different quantities of G at any single moment, and where individuals will have enjoyed vastly different total quantities of G by the end of their lives. To make this more precise, imagine that there are two classes, and that each member of the rich class enjoys at every moment 250 times as much G as any member of the poor class. W_3 is an unequal world on both the simultaneous segments view and the complete lives view.[34] Because it is a highly unequal world on the complete lives view, any moment will contribute to a highly unequal overall outcome. Hence, according to actualist teleological egalitarianism, every moment has negative distribution value. And because at every moment the members of one class enjoy a far greater quantity of G than the members of the other, every moment also has negative distribution value according to nonteleological egalitarianism.[35]

I suspect that many egalitarians who intuitively "see" intrinsic disvalue in inequalities would hold that W_1 is better than W_2, and that W_2 is, in turn, better than W_3. Moreover, and crucially, I further suspect that many egalitarians would hold that every moment in W_1 is better than any moment in W_2, and that every moment in W_2 is better than any moment in W_3. Yet it is difficult to see how such a ranking and such an evaluation can be made unless the value that nonteleological egalitarianism focuses upon and the value focused upon by actualist teleological egalitarianism are *both* salient.[36] For W_1 is not better than W_2 if one is an actualist teleological egalitarian (for the distributional outcome is the same in both worlds), but W_1 is better than W_2 if one is a nonteleological egalitarian who subscribes to the simultaneous segments view (for there is never an inequality in W_1, while every moment in W_2 displays an inequality). And each moment in W_2 is better than each moment in W_3 if one is an actualist teleological egalitarian (because each moment in W_2 contributes to an overall equality of positive distribution value, thereby conferring a positive value on each moment in W_2), but no moment in W_2 is better than any moment in W_3 if one is a nonteleological egalitarian subscribing to the simultaneous segments view (because each moment in W_2 contains just as great an inequality as does each moment in W_3). Yet if any egalitarian actually does hold that every moment in W_1 is better than any moment in W_2, and that every moment in W_2 is better than any moment in W_3, then she cannot be exclusively an actualist teleological egalitarian or exclusively a nonteleological egalitarian. Rather, she would have to value both of what each of those views considers to be of value.

---

[34] If G were income, then it might seem excessive to imagine one class receiving 250 times that of another. In fact, it has been calculated that the top quintile of our own world's population earns 270 times that of the bottom quintile. See Thomas Pogge, *World Poverty and Human Rights* (Malden, MA: Polity, 2002), p. 233.

[35] These three possible worlds are considerably more unequal versions of the three worlds described in Huemer, *op. cit.*

[36] See Carter, "A Defense of Egalitarianism," *Philosophical Studies*, cxxxii, 2 (2006): 269–302.

CENTRAL LI
UNIVERSI
OF

To make this clearer, any moment in W_1 is better than any moment in W_2 on the basis of the value focused upon by nonteleological egalitarianism, while any moment in W_1 is equal in value to any moment in W_2 on the basis of the value focused upon by actualist teleological egalitarianism. But any moment in W_2 is better than any moment in W_3 on the basis of the value focused upon by actualist teleological egalitarianism, while any moment in W_2 is equal in value to any moment in W_3 on the basis of the value focused upon by nonteleological egalitarianism. Furthermore, on the basis of neither the value focused upon by nonteleological egalitarianism nor the value focused upon by actualist teleological egalitarianism is any moment in W_3 better than any moment in W_2, or any moment in W_2 better than any moment in W_1, or any moment in W_3 better than any moment in W_1. We might consider this parallel: A is bigger than B in height, while B is equal to A in length. But B is bigger than C in length, while C is equal to B in height. Furthermore, in neither height nor length is C bigger than B, or B bigger than A, or C bigger than A. Clearly, A is bigger than B, and B is bigger than C, but only because both height and length determine overall size. By parallel reasoning, any moment in W_1 is better than any moment in W_2, and any moment in W_2 is better than any moment in W_3, only if the value focused upon by actualist teleological egalitarianism *and* the value focused upon by nonteleological egalitarianism both contribute to overall value.[37]

Hence, if many egalitarians do hold that any moment in W_1 is better than any moment in W_2, and that any moment in W_2 is better than any moment in W_3, this strongly suggests that many egalitarians are, in fact, more pluralist than they traditionally have been presumed to be. And this would further suggest that egalitarianism might well be more pluralist than traditionally has been presumed.

VIII

Now, to deal with one possible objection, an egalitarian might switch inadvertently from nonteleological egalitarianism, when regarding any moment in W_1 as better than any moment in W_2, to teleological egalitarianism, when regarding any moment in W_2 as better than any

---

[37] Indeed, the seemingly counterintuitive aspects of the complete lives view (see McKerlie, *op. cit.*) seem to arise because of the value focused upon by the simultaneous segments view, while the seemingly counterintuitive aspects of the simultaneous segments view seem to arise because of the value focused upon by the complete lives view. This suggests that both values count. Perhaps the way to deal with this problem is not to abandon either value, but rather to ascertain how best to trade them off. Again, see Carter, "Towards a Multidimensional, Environmentalist Ethic," "Some Groundwork for a Multidimensional Axiology," and "Value-Pluralist Egalitarianism" for discussions of how value trade-offs might be modeled.

moment in W_3. And this might, *prima facie*, seem incoherent. But it is not incoherent for an egalitarian to value both what nonteleological egalitarians and what actualist teleological egalitarians might each focus upon: namely, the value of equality within simultaneous segments and the value of equal complete lives, respectively, especially when the latter arises as a result of inevitable synchronic inequalities that could be compensated for later. To think that this is incoherent, it might be surmised, is to remain trapped in the grip of the presumption of value monism.

In order to demonstrate that there is no necessary incoherence here, consider the three objects, *A*, *B*, and *C*, mentioned in the previous section. If *A* and *B* are both five feet long, but *A* is ten feet tall while *B* is eight feet tall, then *A* is larger than *B*. And if *C* is, like *B*, eight feet tall, but, unlike *B*, is three feet long, then *B* is larger than *C*. Moreover, *A* is clearly larger than *C*. When we compare *A* and *B*, the difference in height becomes salient; whereas when we compare *B* and *C*, the difference in length becomes salient. Clearly, there is no incoherence in shifting focus from the inequality in height to the inequality in length. Similarly, when comparing any moment in W_1 with any moment in W_2, the nonteleological inequality in the latter is salient. And when comparing any moment in W_2 with any moment in W_3, the teleological inequality in the latter is what becomes salient. Why should this be any less coherent than when comparing the unequal sizes of the three objects, *A*, *B*, and *C*? If all the inequalities in W_2 consist in inevitable temporary inequalities that can be compensated for later, then looking ahead to the overall equality that holds among the complete lives of those living in W_2 easily could lead an egalitarian evaluator to judge any moment in W_2 to be better than any moment in W_3. The teleological inequality therefore will become legitimately salient. But as W_2 displays inequalities within simultaneous segments that are not displayed in W_1, and as W_1 displays no greater inequality with respect to complete lives than does W_2, then whenever an egalitarian evaluator compares W_1 with W_2, the nonteleological inequality displayed in the latter will become legitimately salient.

Thus, if a nonteleological inequality and a teleological equality can both be "seen" within the same world (W_2), then surely there is no incoherence in an egalitarian deploying nonteleological egalitarian considerations when holding any moment in W_1 to be better than any moment in W_2 and then deploying teleological egalitarian considerations when holding any moment in W_2 to be better than any moment in W_3. What would be seriously confused would be to employ teleological egalitarian considerations in order to judge any moment in W_1 to be no better than any moment in W_2, then to employ nonteleological egalitarian considerations in order to judge any moment in W_2 to be no better than any moment in W_3, and then to conclude that no moment in W_1 is any better than any moment in W_3.[29] This would be like arguing that because *A* is no bigger than *B* with respect to length, and because *B* is no bigger than *C* with respect to height, then *A* is no bigger than *C*. In other words, focusing upon the inequalities, rather than the equalities, is what enables one to judge that *A* is bigger than *B* and that *B* is bigger than *C*. Consequently, the difference in height becomes salient when comparing *A* with *B*, and the difference in length becomes salient when comparing *B* with *C*. And this is a perfectly legitimate way to proceed. Likewise, the nonteleological inequality becomes salient when comparing W_1 with W_2, and the teleological inequality becomes salient when comparing W_2 with W_3. And this is no less a legitimate way to proceed.

So, by way of conclusion: we have seen that there is reason to regard egalitarians as even more pluralistic with respect to the values they recognize than has, thus far, been presupposed to be the case. And this is so irrespective of whether egalitarians are themselves conscious of their value pluralism.[30]

ALAN CARTER

University of Glasgow

---

[29] This, it can be argued, is precisely the error that Huemer commits in his critique of egalitarianism. See Huemer, *op. cit.*; for a rebuttal, see Carter, "A Defense of Egalitarianism."

[30] It seems to me that several objections to egalitarianism can be argued to confuse deontic injunctions with axiology—a confusion that I hope to address in future work. The focus of the present article has been exclusively axiological, not deontic.

CONCEPTS AND ABILITIES IN ANTI-INDIVIDUALISM[*]

This paper is prompted by two realizations. The first realization is that most critical studies of Tyler Burge's "Individualism and the Mental" focus on the first step of the famous arthritis thought experiment, in particular on the claim that people can be said to think with concepts that they incompletely understand in significant and systematic ways.[1] Less attention by far is given to the thought experiment's third and final step, in which it is argued that a counterfactual subject who lacked the concept *arthritis* would be incapable of having any thoughts or other attitudes about "arthritis as such."[2]

No doubt, the notion of incomplete concept mastery invoked at Step 1 of the thought experiment raises a number of rich and complex

---

[*] A distant ancestor of this paper received helpful comments from John McDowell, Thomas Ricketts, Kieran Setiya, James Bogen, J. E. McGuire, Christine Young, Graham Hubbs, and Jo-Jo Koo. Special thanks must go to Christopher Frey, whose extensive and detailed criticisms encouraged me to abandon that first draft and seek new ways of framing the argument, and to Holly Andersen, for continuous and consistently helpful feedback throughout the process For specific comments on a more recent draft, I owe thanks to Gurpreet Rattan, to colleagues at the Centre for the Study of Mind in Nature (where I presented a version of this paper in November 2009), and, not least, to the editors of this JOURNAL, whose eminently perceptive and constructive feedback helped set the final revisions on the right track. This paper was written while on a grant from the Research Council of Norway.

[1] See, for instance, Donald Davidson, "Knowing One's Own Mind" and "Epistemology Externalized," both reprinted in *Subjective, Intersubjective, Objective* (New York: Oxford, 2001); Akeel Bilgrami, *Belief and Meaning* (Malden, MA: Blackwell, 1992), chapters 2–3; Kent Bach, *Thought and Reference* (New York: Oxford, 1987), chapter 13; Gabriel Segal, *A Slim Book about Narrow Content* (Cambridge: MIT, 2000), chapter 3; Tim Crane, "All the Difference in the World," *Philosophical Quarterly*, XLI, 162 (January 1991): 1–25; Jerry Fodor, "Cognitive Science and the Twin-Earth Problem," *Notre Dame Journal of Formal Logic*, XXIII, 2 (April 1982) 98–118; Gary Ebbs, *Rule-Following and Realism* (Cambridge: Harvard, 1997), chapter 8, and *Truth and Words* (New York: Oxford, 2009), pp. 148–51; Reinaldo Elugardo, "Burge on Content," *Philosophy and Phenomenological Research*, LIII, 2 (June 1993): 367–84; Sanford Goldberg, "Do Anti-Individualistic Construals of Propositional Attitudes Capture the Agent's Conceptions?" *Noûs*, XXXVI, 4 (December 2002): 597–621; Åsa Maria Wikforss, "Social Externalism and Conceptual Errors," *Philosophical Quarterly*, LI, 203 (April 2001): 217–31, and "Externalism and Incomplete Understanding," *Philosophical Quarterly*, LIV, 215 (April 2004): 287–94; Gurpreet Rattan, "Intellect and Concept," *The Baltic International Yearbook of Cognition, Logic, and Communication: Meaning, Understanding, and Knowledge*, v (October 2010): 1–61

[2] Tyler Burge, "Postscript to 'Individualism and the Mental'," in *Foundations of Mind* (New York: Oxford, 2007), pp. 151–81, at p. 162. On this, critics appear to follow Burge's own lead: the most substantial discussions by far in "Individualism and the Mental" (reprinted in *Foundations of Mind*, pp. 100–50) are devoted to defeating possible objections to the doctrine of incomplete concept mastery (see pp. 111–14, 116–32). Comparatively little space is set aside for defending the conclusion at Step 3 of the thought experiment.

0022-362X/11/0810/555–575

philosophical questions, rewarding close attention. Nonetheless, this comparative dearth of attention to Step 3 is remarkable, since it is ostensibly here, and not at Step 1, that Burge intends to make the case for the general thesis of anti-individualism: the distinctively metaphysical doctrine that the nature or identity of many of our mental states "depend in a constitutive way on relations that the individual bears to a wider social environment."[8]

In brief, most commentators seem to assume that the real work in the anti-individualist thought experiments is done by the doctrine of incomplete concept mastery at Step 1, so that the conclusion at Step 3 would follow if only Step 1 were granted. The second realization prompting this paper, then, is that this is not true. Thus, the primary aim of this paper is to show that the conclusion offered at Step 3 of the thought experiment will lack warrant whether or not we accept Burge's invocation of incomplete concept mastery at Step 1. Burge's argument cannot secure the conclusion it promises to secure. Moreover, even if it did, this conclusion would have none of the exciting consequences it is generally taken to have.

The official or generally accepted story concerning the anti-individualist thought experiment purports to take us directly from a description of the counterfactual subject's social and physical environment to a conclusion concerning what concepts this subject may have at his disposal, or what kinds of psychological states he is capable of tokening. This conclusion is supposed to carry a certain kind of modal force: it is not just that the counterfactual subject *is* not, as a matter of fact, tokening an *arthritis*-thought on the occasion in question. Rather, given the environing conditions, it would be *metaphysically impossible* for him to think these kinds of thoughts. On the official story, it appears to be largely a matter of indifference whether we pitch the argument in terms of concepts or in terms of cognitive abilities.[4] The one would follow as a trivial consequence of the other, since possession of a concept *C* is construed simply in terms of the subject's ability to think *C*-thoughts, and *vice versa*.[5] Since I will ultimately offer an argument according to which anti-individualism crucially misconstrues the relation between concepts and cognitive abilities, I will treat these matters separately, focusing on concepts first and cognitive abilities second.[6]

---

[8] Burge, "Postscript to 'Individualism and the Mental'," p. 151 Many commentators seem to assume, implicitly or explicitly, that all Burge would need in order to establish anti-individualism would be to establish the cogency of the doctrine of incomplete concept mastery. On this view, anti-individualism simply *is* the doctrine of incomplete concept mastery, or else follows as a trivial consequence of it. I believe this is a mistake. For one, the doctrine of incomplete concept mastery is supported by claims concerning deference (for example, that the subject *intends to defer* to experts or to conventionally accepted usage) which are arguably much too hedge-claused to serve as a major premise in an inference to the particular metaphysical conclusion that Burge is after (see, for example, Burge, "Social Anti-Individualism, Objective Reference," reprinted in *Foundations of Mind*, p. 309 and note 6). But also, it seems clear that someone can accept several important forms of incomplete concept mastery and yet still resist the conclusion offered at Step 3 of the argument. On this, see Wikforss, "Self-Knowledge and Knowledge of Content," *Canadian Journal of Philosophy*, xxxviii, 3 (September 2008): 399–424, at p 413.

Thus, after examining Burge's argument for anti-individualism in more detail (section I), the first task of this paper is to show (section II) that there is no sound argument that can carry us, with the requisite modal force, from descriptions of subjects' social and physical environments to conclusions regarding the range of concepts these subjects may have at their disposal. Individuals are capable of having the relevant kinds of concepts, and are thus capable of thinking the relevant kinds of thoughts, even under the environmental restrictions that the anti-individualist thought experiments impose.

This leads us to the paper's second and more ambitious task (section III), which is to show that even if the inference from environments to concepts were sound (or if there were some way to supplement the premises so as to nontrivially yield the desired conclusion), the anti-individualist still would not have secured any relevant or interesting version of the conclusion that this subject thereby would lack the ability to think *C*-thoughts. The sense of inability which we then will have ascribed to the counterfactual subject is a peculiar and inconsequential sense—what, following Aristotle's discussion in *Metaphysics*, Book IX, Chapter 3, I will call a ***merely* *Megarian sense of inability***.[7] The problem is that the Megarian sense of inability is restricted to synchronic applications—applied to the present case, all it can tell us is that we could not consistently find that someone is thinking a *C*-thought *at the same time* as we say that he lacks the *C*-concept.[8] But this has no bearing on what, for instance, he is capable of doing the very next moment. As Aristotle rightly observed, this restricted, synchronic sense of ability and inability is not one that we should hold much stock by. Certainly, it is not a sense of ability and inability which should pull much weight in descriptions of the cognitive lives of rational animals. Proper and substantive ability talk is inherently diachronic; yet, as I will argue, no diachronic ability claims follow from Burge's stipulation that the counterfactual subject lacks the *arthritis*-concept at some time *t*.

---

[4] See Burge, "Phenomenality and Reference: Reply to Loar," in Martin Hahn and Bjørn Ramberg, eds., *Reflections and Replies. Essays on the Philosophy of Tyler Burge* (Cambridge: MIT, 2003), pp. 435–50, at p. 438.

[5] See Burge, "Concepts, Definitions, and Meaning," in *Foundations of Mind*, pp. 291–306, at p. 291; "Social Anti-Individualism, Objective Reference," p. 308.

[6] A small specification: my discussion will center on a subclass of what linguists call "lexical concepts," namely, concepts that denote *social kinds*, for example, *contract*, *symphony*, *democracy*, and (surprisingly to some) *arthritis*. I believe my arguments have significant potential for generalization beyond this class of cases but will not pursue that possibility here. Further, we will follow Burge in thinking of concepts as Fregean *Sinne*, or at least as standing in one-to-one relations with Fregean *Sinne*. Thus, concepts are *modes of presentation* of objects, where "objects" denotes a logical category encompassing events, states, processes, properties, classes of objects, and much else in addition to ordinary physical objects.

[7] Translation by W. D. Ross in Jonathan Barnes, ed., *The Complete Works of Aristotle The Revised Oxford Translation*, vol. 2 (Princeton: University Press, 1985).

In terms of the proper, diachronic sense of ability, it will turn out (section IV) that the conclusion of the anti-individualist thought experiment is simply false: the counterfactual subject has the ability to think *C*-thoughts after all, in spite of the fact that he lacks the *C*-concept. This insight receives further support from reflection on the possibility of concept acquisition. Simply put, we have no way of accounting for the acquisition of such concepts which would not presuppose the very cognitive abilities that possession of the concept was supposed to explain. In addition to further undermining Burge's argument for anti-individualism, this realization forces (at least) two important emendations to the view that concepts are abilities, which I explore in section V. First, if concepts are abilities, they are a *sui generis* form of ability and should not be assimilated to standard paradigms of such. Much of the present confusion stems from just such overhasty assimilations, particularly concerning the subject's acquisition and first exercise of these abilities. Second, possession of a concept *C* cannot simply be identified with having the ability to think *C*-thoughts (since the counterfactual subject turns out to have the ability even though he does not have the concept); rather, it should be construed as a distinctive (and important) *way* of having that ability.

I. INTRODUCING ANTI-INDIVIDUALISM

Burge defines anti-individualism as the view that there exist relations of constitutive dependency between the kinds of thoughts a person can think—or types of representational states he can token—and his wider social environment. Absent the relevant environing conditions, it would be impossible for a given person to have certain kinds of concepts, and accordingly, impossible for him to think thoughts containing those concepts.

---

[8] This follows trivially from the implied definition of what it is to have a concept: having a concept *C just is* having the ability to think *C*-thoughts, and *vice versa*.

The main pivot of Burge's argument to this conclusion is a thought experiment proceeding in three steps. In the first step, we are to imagine an otherwise competent speaker of English who utters the sentence "I have arthritis in my thigh." We are to ask, to which attitude is this person giving expression? The thought experiment is designed to elicit from us the following knee-jerk reaction: it could not be a thought involving our concept *arthritis*, since this concept, by definition, applies only to ailments in the joints. It is assumed that someone who in such a way fails to know the meaning of the word 'arthritis' (or in object-level terms, someone who does not know what arthritis is) could not be said to possess the concept *arthritis*. Accordingly, we would be wrong to ascribe to him a thought containing that concept.[9]

However, this knee-jerk reaction is supposed to crumble in the face of deeper reflection on the case. Our initial inclination belies a peculiar blind spot in our thinking about these matters, which Burge's argument aims to isolate and contain. This is the idea that *we can only think with concepts we completely understand*.[10] According to Burge, reflection on the finitude of our cognitive powers suggests that, quite probably, very few of us fully master in this sense any significant range of the concepts which nonetheless appear to be at our cognitive disposal. If our thinking were constrained to operate with concepts which we completely understand, then very little thinking would ever take place. In particular, the idea that we can learn from each other and acquire knowledge by testimony appears to require that we be able to say that the doctor and his patient may share the concept *arthritis*, even though the patient's grasp of this concept is importantly deficient or incomplete.[11] If this doctrine of incomplete concept mastery is cogent, then the main obstacle to our asserting that the subject of our thought experiment is thinking with our concept *arthritis* would seem simply to fall away.

---

[9] In other words, this reaction assumes that what Timothy Williamson calls the "paradigmatic way" to possess a concept is in fact the only way to possess a concept. See Williamson, "Blind Reasoning II: Understanding and Inference," *Aristotelian Society Supplementary Volume*, LXXVII, 1 (2003) 249–93, at p. 250: "As concept possession is usually conceived, a paradigmatic way to have a concept C is to understand a word that means C."

[10] See, in particular, Burge, "Individualism and the Mental," p. 131. To forestall confusion, I ask the reader to note that where I speak of "concepts," "Individualism and the Mental" frequently speaks of "notions" (defined in *ibid.*, p. 102). Burge later switches back to the more widespread nomenclature.

[11] See, for instance, *ibid.*, pp 123, 125–26, 149. For more recent and thorough expositions of this kind of view, see Sanford Goldberg, *Anti-Individualism: Mind and Language, Knowledge and Justification* (New York: Cambridge, 2007); and Ebbs, *Truth and Words*, chapter 5.

In Step 2 of the thought experiment we are to imagine an individual who is identical to the subject of Step 1 in all respects but one: this individual belongs to a speech community which classifies rheumatoid ailments differently than does English. There is no word in his language which covers all and only the diseases which are covered by the English word 'arthritis'. In this counterfactual language, the word 'arthritis' denotes all rheumatoid ailments. Thus, when this individual utters the syntactic string "I have arthritis in my thigh," he, unlike his counterpart in our community, is saying something which may or may not be true.

Finally, in Step 3 of the thought experiment we are to ask how, more specifically, we should understand what this counterfactual person is saying, which thought he is expressing with this phrase. Burge's conclusion is striking: however we are to understand him, we are *not* to understand him as saying or thinking anything that would involve our concept *arthritis*. The reason is simple: he does not have this concept.[12] Burge writes:

> In the counterfactual situation, the patient lacks some—probably *all*—of the attitudes commonly [that is, actually] attributed with content clauses containing 'arthritis' in oblique occurrence. He lacks the occurrent thoughts or beliefs that he has arthritis in the thigh, that he has had arthritis for years, that stiffening joints and various sorts of aches are symptoms of arthritis, that his father had arthritis, and so on.[13]

In short, all *arthritis*-thoughts are out of cognitive reach of the counterfactual subject, because he lacks the conceptual resources to form such thoughts. He lacks these conceptual resources because there is no such concept in currency in his social environment.[14]

---

[12] In short, since 'arthritis' in our language and 'arthritis' in the counterfactual language do not denote the same range of objects, they are not even candidates for expressing the same concept. See Burge, "Individualism and the Mental," p. 106.

[13] *Ibid.*

[14] In this sense, Burge's thought experiment is often taken to complement the *natural-kind externalism* familiar from Hilary Putnam's "The Meaning of 'Meaning'," reprinted in *Mind, Language, and Reality: Philosophical Papers*, vol. 2 (New York: Cambridge, 1975), with a parallel thesis about *social-kind externalism*. On a standard retelling, the structure of each argument is the same, as would be the force of the conclusion obtained through each argument. I believe this assessment overlooks a fundamental difference between the two arguments, which turns precisely on the role allotted to concepts in each. For it is not their lack of a concept *water* that explains why Twin-Earthlings are unable to think *water*-thoughts, but rather the lack of the substance water in their natural environment to serve as the *res* of their *de re* thoughts. By contrast, there certainly are cases of arthritis in Burge's counterfactual community (indeed, for all we know, the counterfactual subject might even be afflicted). What explains the counterfactual subject's alleged inability to think *arthritis*-thoughts, then, is not the absence of arthritis in his natural environment, but rather the absence of the concept *arthritis* in his social environment. Thus, Burge's argument depends on a particular view of the relation between concepts and abilities of which Putnam's argument is entirely innocent. For

If this conclusion is warranted, it provides what Burge needs to establish the truth of anti-individualism. The thought experiment is designed so that ultimately only the differences in the social environment—in the case at hand, practices of medical classification and the linguistic and epistemic norms they give rise to—could explain why the two individuals are expressing different thought contents when each utters the syntactic string "I have arthritis in my thigh." This would be sufficient to establish the claim that there are relations of constitutive dependency between the kinds of thoughts an individual can token and his social environment, thus establishing anti-individualism. The counterfactual subject lacks the concept *arthritis* because there is no such concept in currency in his social environment. Lacking the concept, he is incapable of thinking any such thoughts.

But right away, we notice a hint of hesitation in Burge's argument concerning exactly what the force of the conclusion is supposed to be. On the one hand, some claims in "Individualism and the Mental" seem to suggest a weaker reading, for instance:

> I also do not claim that the fact that our subject lacks the relevant beliefs in the third step follows from the facts that I have described. The point is that it is plausible, and certainly possible, that he would lack those beliefs.[15]

On the other hand, a slew of later statements clearly favor a much stronger reading. For instance:

> The arguments center on the point that in the original situation an individual has one set of thoughts, and in the counterfactual situation the individual *cannot have those same thoughts.*[16]

> The fundamental reasoning [in "Individualism and the Mental"] concerns *conditions under which one can be in certain sorts of mental states, or have certain concepts.* The intuitions on which the thought experiments rely center on *conditions under which it is possible or impossible to have certain thoughts or perceptions.*[17]

---

a recent example of this conflation in the literature, see Williamson, "Sosa on Abilities, Concepts, and Externalism," in John Greco, ed., *Ernest Sosa and His Critics* (Malden, MA: Blackwell, 2004), pp. 236–71, at pp. 263–64. "Believing that one is near water is an externalist mental feature. Presumably, Twin-Otto fails to believe that he is near water because, unlike Otto, he lacks the concept *water.*" If we wish, we can certainly describe Putnam's problem in terms of concepts, so long as we bear in mind that the reference to concepts pulls no real explanatory weight in the thought experiment (which is why Williamson's "because" seems so misplaced). By contrast, the reference to concepts pulls *all* of the explanatory weight in Burge's thought experiment.

[15] Burge, "Individualism and the Mental," p. 117

[16] Burge, "Postscript to 'Individualism and the Mental'," p. 156, my emphasis.

[17] *Ibid.*, p. 162, my emphases.

> The third stage [of the thought experiment] indicates that in such a situation it is *not possible for the individual to have thoughts about arthritis as such.*[18]

> The main case [for anti-individualism] is a set of thought experiments that show that a given person can, under certain circumstances, have a given thought or attitude; but *if certain environmental conditions were different or lacking, a counterpart person could not, as a matter of metaphysical necessity, have that same thought or attitude.* The point can be seen in terms of concept possession: Given certain background conditions, the individual on earth can have a concept *aluminum* or *arthritis*...and the relevant individual on Twin Earth cannot.[19]

The point is not, then, that the counterfactual subject is not, *as a matter of fact*, thinking *arthritis*-thoughts on the occasion in question. Rather, the point is that, given how things stand in his social environment, he *could not be* in possession of the *arthritis*-concept or be thinking any *arthritis*-thoughts. It should be clear that anti-individualism as a metaphysical thesis requires this stronger, modally inflected reading of the conclusion.

But even assuming that reading, what is being claimed remains ambiguous. What makes for the impossibility? The last quote above seems clear: absent certain environing conditions, it would be impossible for the counterfactual subject to think certain kinds of thoughts. "Individualism and the Mental" aims to showcase "the role of social institutions in shaping the individual and the content of his thought."[20] In particular, the argument centers on semantic conventions and the epistemic norms they encode. Sometimes, however, Burge pitches the argument directly in terms of concepts. Yet these are, we may surmise, not competing forms of explanation, but rather two elements of a single line of thought: outside of a certain kind of environment, an individual could not have a certain concept. Without the concept, he could not think the relevant kinds of thoughts.

Thus, it would appear that a fully fleshed-out version of the underlying reasoning would do well to distinguish three elements:

Figure 1

[a] Conceptual resources of public language (or social environment)

[b] → [Determine] → Conceptual resources of individual

[c] → [Determine] → Range of thoughts individual is capable of thinking

Our task is now to assess the cogency and plausibility of this general model.

---

[18] *Ibid.*, my emphasis.

[19] Burge, "Phenomenality and Reference: Reply to Loar," p. 438, first emphasis mine.

[20] Burge, "Individualism and the Mental," p. 100.

II. THE INFERENCE FROM ENVIRONMENT TO CONCEPTS

Recall that we are looking for a way to ground anti-individualism's distinctive metaphysical thesis. To that end, let us look first at the transition from [a] to [b] in Figure 1 above. There are two ways one might try to cash out this transition: either (1) in terms of the semantic conventions and lexical structure of the counterfactual language, or (2) in terms of the social epistemic practices that underlie these semantic conventions and whose norms these conventions encode when all goes well. Both ways of construing the argument can claim textual support. Regarding (1), we note that Burge's primary source of data is that the term 'arthritis' in the counterfactual language has a different meaning than its English homophone.[21] Moreover, Burge is explicitly concerned to add stipulations to the effect that no other word in the counterfactual language means the same as 'arthritis'.[22] Regarding (2), Burge stipulates that in the counterfactual environment, "arthritis has not been singled out for special mention" in medical theory or practice.[23] Presumably, this is intended to specify that even though the counterfactual subject may entertain thoughts *about* arthritis (*de re*), these will not be thoughts about arthritis "as such."[24] Obviously, the two possible routes to the conclusion are often intermingled: if arthritis had been singled out for "special mention," it presumably would have been with the use of a term that would be at least co-extensional with the English term 'arthritis'.

As I shall argue, however, neither (1) nor (2) can secure what Burge's argument requires, namely, a metaphysically necessary inference from social environment to the range of concepts a subject may have at his disposal, and so, further down the line, to the range of thoughts that he is capable of thinking.

Let us look briefly at each, starting with (1), the option that focuses on semantic conventions and the lexical structure of the language. Once this view is made explicit, it can be dismissed rather swiftly by pointing out that it would commit us to the hyper-Whorfian claim that a person could have only such concepts as there are words to express in his public language.[25] Whatever one's views on the relation of language to thought more generally,[26] it would be an absurd view of concepts which pinned their existence—or at least their availability to cognizers—on lexicalization in a public language. Such a view would, for instance, rule out *a priori* that speakers of different languages may share thoughts in areas where the lexical structures of their languages are not completely isomorphic. Arguably, the situation might be even worse: Burge stipulates that the counterfactual community has no word that matches the *extension* of the English word 'arthritis'. But it does not follow that a term co-extensional with 'arthritis' would express the concept *arthritis*. In fact, by Burge's own Fregean lights, it would seem that a construction such as 'rheumatoid inflammation of the joints' (supposing it were co-extensional with 'arthritis') would no more express the same concept as 'arthritis' than 'water' expresses the same concept as 'H_2O'. This hyper-Whorfian view thus would entail that countless banal truths of English would be not just inexpressible in any other language, but unthinkable to monolingual speakers of those languages.[27]

---

[21] See, for instance, *ibid.*, pp. 105, 106.

[22] *Ibid.*, p 106.

[23] *Ibid.*, p. 112; see also p. 104.

[24] See, for instance, Burge, "Two Thought Experiments Reviewed," in *Foundations of Mind*, pp. 182–91, at pp. 184, 189.

[25] See Benjamin Lee Whorf, "Science and Linguistics," in John B. Carroll, ed., *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (Cambridge: MIT, 1956). This view would warrant the epithet "hyper-Whorfian" in being significantly stronger than anything that could reasonably be attributed to Whorf himself.

Such are the easily ascertainable consequences of the view that differences in the lexical structure of languages entail differences in the range of concepts available to speakers of those languages. We can only speculate as to why so many philosophers are still drawn to this way of stating the thought experiment. One possibility is that the methodology of thought experiments itself lets us down here. Maintaining an *a priori* conviction of this sort comes at low cost when we are considering merely hypothetical examples such as Burge's counterfactual community. The conviction is considerably harder to maintain once we realize how many actual languages mirror Burge's counterfactual language in this regard; for instance, several Germanic languages, such as German, Swedish, Norwegian, and Danish, appear to have no term for arthritis as such. It is wildly implausible to hold that this fact alone entails the conclusion that monolingual speakers of these languages must lack the concept *arthritis*. Reflecting on real-world examples of lexical incongruence between languages should make us much less confident of the point at stake in Burge's thought experiment, namely, that the counterfactual subject would lack some cognitive ability which the actual subject has merely in virtue of the different lexemes available to them in their public languages. The counterfactual subject might lack a particular semantic ability—that is, the ability to utter sentences expressing propositions that take *arthritis* as a conceptual component—but we would be wrong to infer from this a relevant difference in cognitive ability.

---

[26] Consider, for instance, Michael Dummett's controversial view that analytic philosophy is committed to the belief that "first, a philosophical account of thought can be attained through a philosophical account of language, and secondly, that a comprehensive account can only be so obtained." Dummett, *Origins of Analytical Philosophy* (Cambridge: Harvard, 1993), p. 4. There is nothing in this view that would support or even suggest the supposition that a person can only have such concepts as there are words to express in his language.

[27] This is a good place to remind ourselves just how few co-extensional terms there are between any two natural languages, let alone how few that would express the same Fregean concept. See, for instance, Kai von Fintel and Lisa Matthewson, "Universals in Semantics," *The Linguistic Review*, xxv, 1–2 (April 2008): 139–201, which claims to find "remarkably few convincing semantic universals" (p. 139) at the lexical level, even as they limit their consideration to candidates for co-extensionality.

Thus, we can reject the view that facts about the lexical structure of the counterfactual language entail that the counterfactual subject would be unable to think *arthritis*-thoughts, let alone that it would be metaphysically impossible for him to do so. Moreover, we note that Burge more recently has sought to distance anti-individualism from this line of thinking, explicitly recanting any strong reliance on suppositions concerning the relation of conventional language to thought.[28]

This leads us to the second way of attempting to cash out the inference from environment to concepts. For what really matters, we may suppose, are not semantic conventions and lexical structure *per se*, but rather the social epistemic practices that underlie these conventions, whose norms these conventions embody and express when all goes well. Looking at the lexical structure of the language might constitute a first-pass epistemic indicator of what concepts are in currency in those practices, but it is no more than that.

But even if we change our focus from linguistic conventions to the epistemic practices that underlie these conventions, we still have no plausible argument to take us directly from environing conditions to the kinds of concepts that a person may have. While this approach avoids the obvious flaws of the linguistic route, it does not manage to shake off the underlying problem. For clearly, someone is capable of having concepts other than those that have currency in the social epistemic practices that surround him.[29] If he is capable of having these concepts in the absence of these environing conditions, then he is capable of thinking the kinds of thoughts in question.[30] The contrary view is plainly undermined by the fact that every social-kind concept we have has arisen at some point in history from a social epistemic practice that did not have that concept. We have no way of accounting for the historical emergence of these concepts other than by conceding that individuals are capable of developing (and thus of having) concepts other than those that have currency in their social environment.

---

[28] See Burge, "Intellectual Norms and Foundations of Mind," in *Foundations of Mind*, pp. 254–74; "Postscript to 'Belief *De Re*'," in *ibid*, pp. 65–81, at pp. 65–66; "Postscript to 'Individualism and the Mental'," pp. 157–62; "Phenomenality and Reference: Reply to Loar," p. 449n3; "Reply to Christopher Gauker," in Maria J. Frapolli and Esther Romero, eds., *Meaning, Basic Self-Knowledge, and Mind: Essays on Tyler Burge* (Stanford: CSLI, 2003), pp. 246, 250.

[29] Several key arguments in Burge's later article "Intellectual Norms and Foundations of Mind" presuppose exactly this point.

In fact, the historical emergence of our concept *arthritis* provides a perfect case in point. This concept developed out of a predecessor concept which, like the concept in currency in Burge's counterfactual community, extended to all rheumatoid ailments.[31] In this sense, we could view the counterfactual language community simply as a past stage of our own. Clearly, then, members of that language community must have been capable of developing the concept *arthritis*, since, as a matter of historical fact, they did.

I conclude that the official story behind Burge's anti-individualism has come up short: there is no good argument which would take us directly from environing conditions to concepts with the metaphysical certainty that anti-individualism requires. Emphasis on neither the lexical structure of the language nor the conceptual structure of the underlying epistemic practices will do, since it is plainly possible for the counterfactual subject to have the concept *arthritis* in the absence of these environing factors. Quite simply, Burge cannot make induction into a social practice that already has the concept *arthritis* a condition of possibility for acquiring that concept, on pain of making it impossible to account for how anyone could have come to have the concept in the first place. While it certainly may be true that, ordinarily, most of us acquire our social-kind concepts by the agency of our social peers, such interaction cannot be a metaphysical condition of possibility for acquiring any one of these concepts.[32]

---

[30] Even in the presumably favored case of natural-kind externalism, the inference from environment to cognitive ability is not as straightforward as many seem to suppose. For instance, presuming that Twin Earth scientists have the concepts *hydrogen* and *oxygen*, why should it be impossible for them to develop the concept *H_2O*? For a parallel line of thought in Burge, see "Other Bodies," reprinted in *Foundations of Mind*, pp. 82–99, at p. 98n18.

[31] See Morris Longstreth, *Rheumatism, Gout, and Some Allied Disorders* (New York: William Wood, 1882), p. 1.

[32] I can think of only one way in which one might even try to make such a case, namely, by treating social-kind terms somewhat as Saul Kripke proposes that we treat fictional names. See Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980), pp. 157–58. Thus, one might try to argue that just as no one could have beliefs about Sherlock Holmes unless they were causally connected, at however many stages of removal, to Arthur Conan Doyle, so no one could have beliefs about arthritis as such unless they were causally connected to our language community and the epistemic

III. THE INFERENCE FROM CONCEPTS TO ABILITIES

So far, then, we can conclude that Burge's official story, which claims to take us directly from a description of the counterfactual subject's social environment to a conclusion about the range of concepts that he may have at his disposal, cannot be made good. No matter how empirically robust such a connection might be, it will remain a contingent connection. As such, it cannot support the counterfactuals that Burge's conclusion requires.

But we should not stop here. The fact that we have blocked the transition from [a] to [b] in Figure 1 above does not mean that we should not have an interest in assessing the transition from [b] to [c] in its own right. For instance, someone might concede the arguments of the previous section and still hold that if (contrary to fact) the argument were sound, then the conclusion would be interesting and relevant in precisely the way that anti-individualists contend. Thus, one might suppose that if we had some way of supplementing the premises so as to pave the way for a valid inference to the proposed conclusion, then anti-individualists would have what they need.

In light of this, we should have an interest in assessing the conclusion and its consequences on its own terms. As we shall see, a separate and in many ways deeper problem attaches to the proposed inference from [b] to [c] and to the conclusion that such an inference would allow us to obtain. Even though, on the face of it, the thought experiment issues in a claim about abilities, it will turn out to be a peculiar and inconsequential sense of ability that is at stake. Reflection reveals that this is not a sense of ability which we should take as being relevant to our descriptions of subjects' cognitive lives more broadly. Thus, even if the argument were sound, its conclusion would not have the exciting consequences it is generally taken to have.

The only kind of ability claim which can be drawn from an argument of this general form is what, following Aristotle, we may call *a*

---

practices that underlie it. I doubt that a plausible case can be made for this analogy. What gives the Kripkean treatment of fictional names its plausibility is the fact that Sherlock Holmes does not exist, and so, that no one could have *de re* thoughts about him. By contrast, arthritis does exist, quite independently of any particular conceptual classification. Anyone can have *de re* thoughts about arthritis, as might Burge's counterfactual subject when he wakes one morning and thinks *that is painful*. Treating social-kind terms along the lines of fictional names would rule out *a priori* that different (unconnected) social practices could independently develop the same concept *arthritis* (and likewise for other social-kind concepts). If this were a sound strategy, then Burge's thought experiment would no longer need to stipulate that the counterfactual language term 'arthritis' has a different extension than our homophone. Co-extensional terms in languages that are lexically isomorphic at every point would serve the thought experiment just as well, so long as it was stipulated that there had been no interaction between our language community and the counterfactual community.

merely Megarian ability claim. It is Megarian in the sense that it has only a synchronic application:[33] all the anti-individualist thought experiment could tell us is that we cannot consistently say that someone is thinking a *C*-thought *at the same time* as we say that he lacks the *C*-concept. Construed as an ability claim, however, this conclusion has all the force and cogency of the Megarians' observation that no one who is standing is capable of sitting. No doubt, there is a sense in which this observation is true (on the supposition that sitting and standing cannot be predicated of a single subject at a single time). But proper ability talk, as Aristotle rightly points out, has an inherently diachronic dimension: it has to do with the state or position that a person or thing can occupy next. As I will argue, were we to try to read a diachronic dimension into the conclusion of the arthritis thought experiment, that conclusion would no longer be supported by the argument. In the proper, diachronic sense, the counterfactual subject will have the ability to think *C*-thoughts even though he lacks the *C*-concept.

We can begin by noting that anyone who holds that concepts are abilities presumably will allow that there is a sense in which someone who does not have the concept *C* still may be said to be able to think *C*-thoughts. That is, such a subject still might be *able to acquire* the concept *C*, thereby rendering himself able to think *C*-thoughts. According to Aristotle, "a thing is capable of doing something if there is nothing impossible in its having the actuality of that of which it is said to have the capacity."[34] If this is correct, then we might think that we already have a simple recipe for generating counterexamples to the view that concepts are abilities. For evidently, there is nothing impossible (barring Megarian restrictions) about a subject coming to think *C*-thoughts even though he (now) lacks the *C*-concept.

However, an advocate of the view that concepts are abilities can plausibly retort that Aristotle's dictum (at least as applied here) misses the importance of the distinction between having the ability to Φ and having the ability to acquire the ability to Φ. Thus, to borrow an example from Gilbert Ryle, we might say (truly) of John Doe and Richard Roe alike that he can swim.[35] But these will be relevantly different senses of "can": John Doe can swim in the sense that he has been taught how to swim and has not forgotten. Richard Roe can swim only in the sense that he is able to learn how to swim. Nonetheless, as Ryle observes, "It would be wrong to predict about [Richard Roe], what it would be right to predict about an idiot, that since he now flounders helplessly in the water, he will still flounder helplessly after he has been given tuition."[36] However, even though John Doe and Richard Roe alike satisfy Aristotle's criterion on abilities, anyone taking a clear-eyed view of the situation will be forced to acknowledge that there is a significant difference between the two cases. I propose that we cash out the difference in terms of a distinction between primary and secondary abilities: John Doe has the ability to swim in the primary sense, whereas Richard Roe has the ability merely in the secondary sense (that is, he is able, in the normal run of things, to acquire that ability).

---

[33] Thus, I am *not* claiming that the anti-individualist position on concepts and abilities is Megarian in the (perhaps more familiar) sense according to which a subject *S* has the ability to Φ only at such times as he is actually exercising that ability. For a useful perspective on this distinction, see Stephen Makin, "Megarian Possibilities," *Philosophical Studies*, LXXXIII, 3 (September 1996). 253–76.

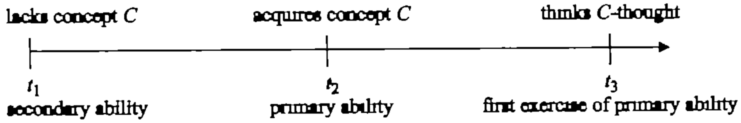
[34] *Metaphysics*, Book IX, Chapter 3, 1047a24–25.

[35] Gilbert Ryle, *The Concept of Mind* (Chicago: University Press, 2002), pp. 128–29

Thus, abilities in the secondary sense are not counterexamples at all, but serve to bring out the essential line of thought animating the view that concepts are abilities: in order to possess the ability to think *C*-thoughts in the primary and proper sense, the person who lacked the concept *C* would precisely first have to acquire that concept. It is presumed throughout the discussion, and thus not a counterexample to such a view, that any rational mature adult may be able to acquire the concept in question, and thus may yet be able to think *C*-thoughts in a secondary sense. The point can be seen in terms of the Fregean dictum that concepts are the constituents of thoughts. Crudely put, if you do not have the concept, then you cannot form any thoughts in which that concept would figure as a constituent. This in no way precludes, however, that you may be able to acquire the concept; then, and only then, will you have the ability to think the kinds of thoughts in question in the proper and primary sense.

This gives us a clue as to how the anti-individualist must construe the diachronic dimension of the claim that concepts are abilities: at some t_1, the subject does not have the concept *C* and thus does not have the ability to think *C*-thoughts. Then, at t_2, he acquires the concept, thereby rendering himself able, at t_3, to think *C*-thoughts. The crucial point is that the subject's acquisition of the concept *C* must precede his first tokening of a *C*-thought, lest the distinction between primary and secondary abilities collapse.

Figure 2



---

[36] *Ibid.*, p. 129.

Plausibly, variations on this schema can cast some light on the acquisition of certain paradigmatic examples of practical abilities.[37] As I will argue, however, it cannot be made good in connection with concepts; in particular, it turns on a notion of *concept acquisition* of which we can make neither empirical nor theoretical sense. Quite simply, we have no way of accounting for a subject's acquisition of a concept without presupposing the very cognitive abilities that possession of the concept was supposed to explain. On the contrary, subjects must be able to transition directly from not having the concept *C* to thinking *C*-thoughts without the interposition of a distinct stage of concept acquisition. On the most favorable reconstruction the anti-individualist can muster, the subject's acquisition of the concept *C* will be coeval with his first tokening of a *C*-thought.[38] It follows that the subject already must have had the ability (and not just in a secondary sense) even though he did not have the concept. Accordingly, we would be wrong simply to identify possession of the concept *C* with the ability to think *C*-thoughts. At best, having the concept *C* will turn out to be a distinctive way of having that ability, but not the only way.

IV. THE PROBLEM OF CONCEPT ACQUISITION

If this line of thought is correct, then it follows that the anti-individualist would be wrong to conclude from a subject's lacking the concept *C* that he would be incapable of thinking *C*-thoughts. That is, the inference from [b] to [c] is invalid, unless the scope of the ability-operator is restricted to synchronic assessment, in which case it would follow trivially, merely as a matter of unpacking the implied definition of what it is to have a concept.

---

[37] I have in mind examples like swimming (Ryle) or house-building (Aristotle). All too briefly, what makes for the difference? Paradigmatic practical abilities such as swimming or house-building are skills which permit of indefinite sharpening, where the cut-off line between mastering the skill and being in the process of acquiring the skill is, while not entirely arbitrary, largely a matter of convention. This leeway permits us to say, at some time *t*, that a subject now has the ability although he has never exercised it. In a real sense, the person who is learning to swim is *swimming* before such time as we shall say that he is a swimmer. However, by convention we call what he does up to that point not an exercise of the ability to swim, but rather steps toward the acquisition of that ability. There is no comparable "period of apprenticeship" for the sorts of cognitive abilities which concepts are supposed to represent. Throughout, we have been assuming, with Burge, that if someone is thinking a *C*-thought, then he has the *C*-concept. (For complications, see next footnote.)

[38] Certainly, plausible reconstructions *less* favorable to anti-individualism are not far to seek. One such reconstruction deserves mention. Here, the acquisition of the concept *C* may lag far behind the subject's first tokening of a *C*-thought. It takes not one but several competent tokenings of *C*-thoughts in order to possess the concept *C*. This possibility is worth pointing out because it threatens to undermine the intuitive basis of Burge's doctrine of incomplete concept mastery. For on this view, we are free to attribute *C*-thoughts to some subject without thereby committing ourselves to saying that he has the *C*-concept, just as we can say (see previous footnote) that someone is swimming—or is performing such motions as otherwise might count as swimming—without thereby saying that he has the ability to swim. On this view, the fact that our subject is tokening a thought such as *I have arthritis in my thigh* is no longer decisive evidence that he has the concept *arthritis*, but rather perhaps, evidence against it.

To see this, we may consider an example: legal scholars point out that there is no term in the French language which expresses the legal concept *self-defense*. The French term closest in meaning to the English 'self-defense' is '*légitime défense*', which applies equally to defense of self and to defense of (nonconsenting) others.[39] Let us take this as an occasion simply to stipulate that at some time t_1, the French did not have the concept *self-defense*, and so, by Burgean hypothesis, were incapable of thinking thoughts about self-defense as such.[40] Yet today they clearly are capable of thinking *self-defense*-thoughts, inasmuch as they are knowing and willing signatories, for instance, to the United Nations Charter, Article 51 of which crucially involves precisely that concept. So our question now is, how might the French have gone from t_1, at which point they lacked the *self-defense*-concept, and so, by hypothesis, were incapable of thinking *self-defense* thoughts, to some t_{1+n}, at which point they were capable of thinking *self-defense*-thoughts, and so, by hypothesis, must have had the *self-defense*-concept?

Once the question is spelled out to this level of detail, it is easy to locate the problem with the temporal schema laid out in Figure 2. For this schema leaves us with no good account of *concept acquisition*, that crucial step which would explain the transition from t_1 to t_3. We have no way of accounting for a subject's acquisition of a concept which would not presuppose the very cognitive abilities that possession of the concept was supposed to explain. Simply put, one cannot acquire a concept except by instantiating it in thought.

Concepts, on Burge's view, are ways of thinking about objects.[41] How, then, might one go about acquiring a concept? As stated, the question might be too broad to permit an illuminating answer. We may fare better by considering specific instances: thus, for instance, one might acquire the concept *self-defense* by learning that attacks upon oneself give one certain legal prerogatives that one does not get from attacks upon anyone else. Similarly, one might acquire the concept *arthritis* by learning that certain occurrences of rheumatism in the joints have etiological peculiarities not shared by other instances of rheumatism.[42] But these are precisely thoughts which embody or express the way of thinking in question. Generally, by the time we should say that someone has acquired the concept *C*, he will already be deeply involved in thinking *C*-thoughts. On the most favorable reconstruction the anti-individualist can muster, the acquisition of a concept *C* will be coeval with the subject's first tokening of a *C*-thought. Evidently, then, the subject was able to think *C*-thoughts at t_1, even though he did not (already) possess the concept *C*. Accordingly, possession of the concept *C* cannot simply be equated with the ability to think *C*-thoughts, nor will lack of the concept entail the corresponding inability, since this would leave us bereft of the means to say how he ever came to token a *C*-thought in the first place.

---

[39] See George P. Fletcher and Jens David Ohlin, *Defending Humanity. When Force Is Justified and Why* (New York: Oxford, 2008), chapter 3. More specifically, the French could certainly apply the concept *légitime défense* to an instance of self-defense, but this is precisely not to think of it *as* self-defense.

[40] Note that this is pure stipulation intended to fix our minds on the relation between concepts and cognitive abilities, and by no means signals a residual adherence to the view discarded above, namely, that a person can only have such concepts as there are words in his language to express. Indeed, as I will argue shortly, the French *now* possess the concept *self-defense* even as they still lack a lexical item which refers to it uniquely

[41] Again, "objects" is in no way restricted to concrete particulars, but also encompasses events, states, processes, properties, classes of objects, and much else besides.

On the picture that is now emerging, a person's acquiring a concept *C* coincides with that person's first coming to think *C*-thoughts. By contrast, it is quite misleading to picture concept acquisition as the acquisition of some pure psychological capacity or module which yet lies unexercised, as Figure 2 invites us to do.[43] But this is the only picture of concept acquisition which would support the conclusion that a subject who has not (yet) acquired the concept *C* does not (yet) have the ability to think *C*-thoughts.

In defense of his conclusion that the counterfactual subject would be unable to think *arthritis*-thoughts, Burge writes that "it is hard to see how [he] could have picked up the notion of arthritis."[44] But there is nothing mysterious about how he might acquire the concept *arthritis*, or about how the French acquired the concept *self-defense*. They simply actualized abilities which they already had by coming to think of certain objects (actions, events, and so on) in certain ways. By contrast, it certainly would be "hard to see" how the counterfactual subject could have acquired the concept *arthritis* if the acquisition of that concept depended on learning a word that means 'arthritis'. If that were the case, it would be *really* hard—indeed, impossible—to see how anyone came to acquire this concept in the first place. To be sure, the mechanisms by which the actual subject and the counterfactual subject first came to acquire the concept *arthritis* might differ. The actual speaker might have acquired it by learning the conventional meaning of a term in the English language. That route is not available to the counterfactual speaker, who might instead have acquired it by reflecting on the etiological peculiarities of occurrences of rheumatism in the joints. But in neither case would it be correct to say that the acquisition of the concept preceded the first tokening of the relevant kinds of thoughts. Learning the meanings of words involves learning about things. Even when we acquire concepts by word learning, we do so by instantiating them in thoughts.

---

[42] In view of the doctrine of incomplete concept mastery, of course, there need be no particular set of true beliefs (or inferential dispositions) that someone must acquire in order to have acquired a concept *C*. But this provides no comfort for the anti-individualist. For presumably no one could acquire the concept *C* without acquiring *some* true beliefs about C. Let these true beliefs now be the *C*-thoughts which marked his acquisition of the *C*-concept.

[43] For a related line of thought in Burge, see "Individualism and the Mental," p. 107n1

[44] *Ibid.*, p. 106.

V. TOWARDS A BETTER MODEL OF CONCEPTS AND ABILITIES

One source of worry remains, namely, that this line of argument will entail the loss of any meaningful distinction between having and not having a particular concept.

I believe the worry is unfounded. Even though having the concept *C* cannot simply be equated with having the ability to think *C*-thoughts, it can still rightly be counted as a distinctive and psychologically important *way* of having that ability. To see how, consider two persons *A* and *B*. *A* has a long history of thinking *C*-thoughts, whereas *B* has never tokened a *C*-thought in his life. We find it natural to say that *A* has the concept *C* whereas *B* does not. Nonetheless, as I have argued, we would be wrong to infer that *A* has the ability to think *C*-thoughts whereas *B* does not. Rather, they each possess that ability but in distinctive ways. *A*, we might say, has a *local* (or *modular*) ability, which can manifest itself in the following way: he can conjure *C*-thoughts, for instance, simply by recalling them from semantic memory, and without engaging his rational capacities on a broader scale. Depending on the particularities of the concept in question, he might also have an entry in his cognitive lexicon which expresses that concept uniquely.[45] These are among the ways in which *C*-thoughts are immediately accessible to *A*, facilitated by his previous history with such thoughts. This local (or modular) ability is what I propose that we call *having the concept C*.

But *B*, too, has the ability to think *C*-thoughts, even though he has never exercised that ability. We must acknowledge this, for *B* might simply be *A* at some earlier point in time, and as we have seen, without granting him that ability we would be at a loss to explain how he came to acquire the concept in the first place. So, *A* and *B* alike have the ability to think *C*-thoughts. But the way in which they possess the ability is importantly different. *B*'s ability, we might say, is of a *discursive* sort. He is, we presume, a rational being of mature capacities, possessed of language. He can draw on background knowledge, experience, intuition, imagination, perceptual happenstance, and all the other resources that people draw on when they come up with new ways of thinking about things: now he is thinking *C*-thoughts. *Now* we may say that he has the *C*-concept. But we would be quite wrong to insist that he must have somehow acquired that concept in the steps leading up to his first exercise of that ability.

---

[45] This, though, is by no means a necessary condition, as my argument in section II shows.

Finally, we might think that *A*'s and *B*'s ways of accessing *C*-thoughts are sufficiently different, and that the difference is of sufficient psychological relevance, that we would be warranted in thinking of them as different abilities altogether. Thus, *A* does, after all, have an ability that *B* does not; moreover, the difference is rooted in the fact that *A* has the concept whereas *B* does not. Certainly, this is correct as far as it goes. But it matters how we individuate abilities, and with this move we will have changed the subject of discussion altogether. Throughout, the focal point of our discussion has been subjects' ability to think *C*-thoughts. Both *A* and *B* have this ability, even if they can only exercise it by way of different cognitive mechanisms. If *A*'s mechanism for exercise strikes us as significant enough to warrant being called a distinct ability, then the ability we are now contemplating is no longer correctly specified as the ability to think *C*-thoughts *simpliciter*, but rather as the ability to think *C*-thoughts by way of whatever mechanism we fasten on, for instance, semantic memory. But to point out that *B* lacks *this* ability has none of the consequences that are usually taken to flow from Burge's argument for anti-individualism.[46]

VI. CONCLUSION

In sum, I argued in section II that Burge's official argument, which proposes to carry us with a certain kind of modal force from a description of the counterfactual subject's social environment to a conclusion regarding what concepts he can have at his disposal, is not sound. Emphasis on neither the lexical structure of the counterfactual language nor the conceptual structure of the underlying epistemic practices can secure what Burge's argument requires. It follows that the arthritis thought experiment fails to provide support for anti-individualism as a metaphysical thesis about "conditions under which one can be in certain sorts of mental states, or have certain concepts."[47]

---

[46] Moreover, once we start individuating abilities along these lines, it is suddenly an open question whether we should say that any two people whom we might ordinarily think of as having the same concept *arthritis* (for example, the doctor and the patient suffering from incomplete concept mastery) also have the same cognitive ability. In other words, abilities will now be individuated on a much finer grain than we are capable of individuating lexical concepts. This again would jeopardize the basis of the doctrine of incomplete concept mastery.

In section III, the ante was upped considerably. There I argued that even if the argument were sound, the conclusion would be uninteresting. The conclusion would be warranted only if it were restricted to a synchronic, Megarian reading: we cannot consistently say that the counterfactual subject is thinking *arthritis*-thoughts at the same time as we say that he lacks the *arthritis*-concept. By contrast, once we take a proper, diachronic view of abilities, it no longer follows that someone who lacks the concept *C* will be unable to think *C*-thoughts. This finding is supported by reflection on the possibility of concept acquisition (section IV). We simply have no way of accounting for the acquisition of a concept which would not presuppose the very cognitive abilities that possession of the concept was supposed to explain. Instead, the correct view is that a subject acquires the concept in and through his first exercise of the relevant cognitive ability. Accordingly, the subject must already have possessed that ability even though he did not possess the concept.

Two consequences follow from this argument (section V). First, if concepts are abilities, they are a *sui generis* form of ability and should not be assimilated to standard paradigms of such. Much of the present confusion stems from just such assimilations, particularly concerning the subject's acquisition and first exercise of these abilities. Second, possession of a concept *C* cannot simply be identified with having the ability to think *C*-thoughts, but should rather be construed as a distinctive (and important) *way* of having that ability.

ENDRE BEGBY

Centre for the Study of Mind in Nature,
University of Oslo

---

[47] Burge, "Postscript to 'Individualism and the Mental'," p. 162.

COMMENTS AND CRITICISM

AN INSUBSTANTIAL EXTERNALISM*

In one of his recent articles published in this JOURNAL,[1] Alvin I. Goldman argues that since one must count epistemic rules among the factors that help to fix the justificational status of agents (generally called *J-factors*), not all J-factors are internalist, that is, intrinsic to the agent whose justificational status they help to fix.[2] After all, for an epistemic rule to count as a genuine J-factor, it must be objectively correct and, therefore, "independent of any and all minds."[3] Consequently, it cannot be intrinsic to any particular epistemic agent. In this brief commentary, I will argue that Goldman's argument misunderstands what it takes for epistemic justification to be internalist and, therefore, fails to guarantee his externalist conclusion. In particular, I want to demonstrate that Goldman's argument trivializes the difference between intrinsic and extrinsic properties that lays at the basis of the internalist/externalist debate. I will show that, if sound, simple variations on Goldman's argument could be used to prove the absurd conclusion that all properties are extrinsic. Now, since the intrinsic/extrinsic distinction is fundamental to debates in several areas of philosophy, not only the internalist/externalist debate in epistemology, I conclude that Goldman's argument cannot be sound.

This commentary will proceed as follows. First, I will present Goldman's argument against internalism from section VII of his aforementioned article. Then, I will explain how Goldman's way of presenting the internalist/externalist debate in epistemology is related to the intrinsic/extrinsic debate in the metaphysics of properties. Finally, I will show how a simple variation on Goldman's argument can be used to prove that, for any object a and any of its properties P, a is

---

*Many thanks to Claudia Lorena García and Miguel Ángel Fernández for comments on earlier drafts of this text.

[1] Alvin I. Goldman, "Internalism, Externalism, and the Architecture of Justification," this JOURNAL, CVI, 6 (June 2009): 309–38

[2] The relevant extrinsic/intrinsic distinction here is not the "global" classification of properties, but the "local" classification of ways an individual may have a property. See I. L. Humberstone, "Intrinsic/Extrinsic," *Synthese*, CVIII, 2 (August 1996): 205–67 Carrie Figdor, "Intrinsically/Extrinsically," this JOURNAL, CV, 11 (November 2008): 691–718, prefers to use the terms "extrinsically" and "intrinsically," so in this paper I will speak interchangeably of a property being intrinsic (extrinsic) to an object and the object having the property intrinsically (extrinsically).

[3] Goldman, *op. cit*, p. 317.

0022-362X/11/0810/576–582

extrinsically *P*. As an illustration, I will adapt Goldman's argument to derive the mistaken conclusion that the CD I am currently listening to is extrinsically round. Thus, Goldman's original argument will be proved unsound.

Goldman's original argument is best understood as a *reductio ad absurdum* against internalism. It starts by assuming that every justified doxastic state token is so in virtue of conditions (circumstances or causes) internal to the agent. From this assumption, Goldman derives the necessity of postulating an epistemic norm that links such conditions with the justification they are assumed to help explain. Finally, since this norm both helps explain why the agent is justified (and thus must count as one of its J-factors) and is external to the agent, Goldman concludes that not all J-factors are internal to the agent. Therefore, the internalist hypothesis is false. In fuller detail, Goldman's argument from section VII can be reconstructed like this:

(1) A subject *S* is justified in holding doxastic attitude *D* towards proposition *p* at time *t* either internally or externally.
(2) Assume that *S* is internally justified in holding doxastic attitude *D* towards proposition *p* at time *t*.
(3) From (2), every factor involved in accounting for *S*'s being justified in holding doxastic attitude *D* towards proposition *p* at time *t* must be internal to such subject *S*.
(4) Call *C* the set of conditions, circumstances, or causes that help explain why *S* is justified in holding doxastic attitude *D* towards proposition *p* at time *t*.
(5) From (3) and (4), there has to be some objective strong connection between conditions *C* and *S*'s justification, in virtue of which the former helps explain the latter.
(6) For this connection to be objective and explanatory, its validity cannot depend solely on *S*. There must be one or more correct epistemic J-rules that jointly permit a subject who is (or was) in *C* to form or retain attitude *D* towards *p* at *t*.
(7) Thus, from (5) and (6), the correctness of the relevant J-rules must be external to *S*.
(8) Furthermore, from (5), such J-rules must also be counted among the factors that help explain why *S* is justified in holding doxastic attitude *D* towards proposition *p* at time *t*.
(9) From (7) and (8), there is an external factor that helps explain why *S* is justified in holding doxastic attitude *D* towards proposition *p* at time *t* that is not internal to *S*.
(10) From (9), the negation of (3) follows; that is, not every factor that helps explain why *S* is justified in holding doxastic attitude *D* towards proposition *p* at time *t* is internal to *S*.
(11) From (10), the negation of (2) follows; to wit, *S* is not internally justified in holding doxastic attitude *D* towards proposition *p* at time *t*.

(12) Therefore, from (1) and (11), *S* is externally justified in holding doxastic attitude *D* towards proposition *p* at time *t*.

Goldman considers *accessibilism* and *mentalism* as the two main conceptions of what it means for an event, state of affairs, or condition to be internal. According to accessibilism, an event, condition, or state of affairs *J* is internal to an agent *S* if and only if it is "directly" accessible to *S* at the time it obtains. For mentalism, in contrast, a state of affairs, event, or condition *J* is internal to an agent *S* if and only if it is mental and nonfactive. In both cases, *J* is external if and only if it is not internal.

Notice that, on both the accessibilist and mentalist conceptions, whether an event, condition, or state of affairs *J* is internal or external to an agent *S* depends, amongst other things, on whether *J* is intrinsic or extrinsic to *S*. As a matter of fact, it is a necessary (and perhaps also sufficient) condition for *J* being internal to *S* in the epistemological sense that *J* be intrinsic to *S* in the metaphysical sense.[4] Nonfactive mental states are intrinsic to agents that are in them—that is, for any nonfactive mental state *M*, if an agent *S* is in *M*, then such agent is in *M* independently of any external factors. In other words, from an agent *S* being in a nonfactive mental state *M* one cannot infer anything about *S*'s external world. Also, extrinsic properties are not directly accessible to the agents that have them. This means that any event, condition, or state of affairs that is directly accessible to an epistemic agent must be intrinsic to him. In both cases, therefore, what is internal is also intrinsic.

This means that the debate between internalist and externalist theories of justification in epistemology—like many other philosophical debates—substantially depends on how we draw the line between intrinsic and extrinsic properties. Appeal to this basic distinction has also been used to distinguish between real and "merely Cambridge" change,[5] between the semantic and the pragmatic information conveyed by a linguistic utterance,[6] between intrinsic and extrinsic value,[7] and between narrow and wide mental content,[8] among other philosophically significant distinctions.[9] The distinction has also served to explain the semantic properties of nouns,[10] give a nontrivial formulation of the Identity of Indiscernibles,[11] and make sense of scientific structuralism.[12] In other words, the intrinsic/extrinsic distinction has proved very fruitful in philosophy. Unfortunately, Goldman's argument trivializes it, and in doing so he trivializes the whole internalist/externalist debate in epistemology.

---

[4] This means that *J* must either be an intrinsic property of *S* or supervene on intrinsic properties of *S*. See Brian Weatherson, "Intrinsic vs. Extrinsic Properties," *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta, ed., URL = <http://plato.stanford.edu/archives/fall2008/entries/intrinsic-extrinsic/>.

[5] P. T. Geach, *God and the Soul* (London, UK: Routledge, 1968).

[6] Isidora Stojanovic, "The Semantics/Pragmatics Distinction," *Synthese*, CLXV, 3 (December 2008): 317–19

[7] Michael J. Zimmerman, *The Nature of Intrinsic Value* (Lanham, MD: Rowman and Littlefield, 2001).

[8] Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy*, IV, 1 (September 1979): 73–121.

To show this, I will present and motivate a general argument schema that, if sound, would allow us to show, for any property of any object, that it is extrinsic. Now, for the intrinsic/extrinsic distinction to do the philosophical work it is expected to, it cannot be conceptually true that all properties are extrinsic. Otherwise, the distinction would be vacuous and could do no substantial philosophical work. Thus, any argument that fits the aforementioned schema must be unsound. Since Goldman's argument is exactly of this form, it must also be unsound.

Since extrinsic properties depend on external factors, objects may acquire or lose any of them without losing or acquiring any intrinsic properties. In particular, if an object a has a nonvacuous property P extrinsically,[13] it must be possible for a to become not-P without losing any of its intrinsic properties. Therefore, to show that a property P is intrinsic to an object a, it is enough to find a subset of a's intrinsic properties that together entail and help explain why a is P.[14] Despite the *prima facie* soundness of this method, one may feel tempted to use a counter-argument like the following to show that, despite being entailed by internalist properties, P is still extrinsic to a:

(1) If a is P, then a is P either intrinsically or extrinsically.
(2) Assume that a is P intrinsically.
(3) From (2), every factor involved in accounting for a's being P must also be intrinsic to a.
(4) Call C every such factor involved in a's being P.

---

[9] See Weatherson, *op. cit.*, and Humberstone, *op. cit.*, for other applications.

[10] Franck Lebas and Pierre Cadiot, "La Constitution Extrinsèque du Référent: Présentation," *Langages*, CL, 2 (2003): 3–8.

[11] Peter Forrest, "The Identity of Indiscernibles," *The Stanford Encyclopedia of Philosophy* (Summer 2009 Edition), Edward N. Zalta, ed., URL = <http://plato.stanford.edu/archives/sum2009/entries/identity-indiscernible/>.

[12] J. Ladyman, "What Is Structural Realism?" *Studies in History and Philosophy of Science Part A*, XXIX, 3 (September 1998): 409–24.

[13] I am calling a property "vacuous" if it is necessarily possessed by any entity. For example, I vacuously have the property of *being such that* $3+4=7$. For these properties, the internalist argument I am about to present is unsound. A property is nonvacuous if it is not vacuous.

[14] See Figdor, *op. cit.*

(5) From (3) and (4), there has to be some objective and nomologically strong connection between *C* and *P* in virtue of which *C* helps explain why *a* is *P*.[15]

(6) For the connection between *C* and *P* to have the necessary explanatory power, it must generalize to other cases beyond *a*. It is not enough that *Ca* implies *Pa*; it is also necessary that for any *x*, if *Cx* then *Px*, and that the truth of this later generalization be independent of *a* itself.

(7) Thus, from (5) and (6), the objective and nomological connection between *C* and *P* must be external to *a*.

(8) Furthermore, from (3) and (4), this connection between *C* and *P* is also a factor that helps explain why *a* is *P*.

(9) From (7) and (8), there is a factor that helps explain why *a* is *P* that is not intrinsic to *a*.

(10) From (9), not (3). Not every factor that helps explain why *a* is *P* is intrinsic to *a*.

(11) From (10), not (2). *P* is not an intrinsic property of *a*.

(12) Therefore, from (1) and (11), *P* is an extrinsic property of *a*.

Notice that Goldman's argument perfectly fits this schema. Furthermore, notice that any argument of this form is an *a priori* conceptual *reductio* with no substantial premise of its own; it depends on nothing particular about *a* or *P*, so it can be applied to show of any object and any property[16] that the property is external to the object. In this sense, it trivializes the intrinsic/extrinsic distinction. If it were sound, it could be used to show that, for example, the round shape of the CD I am currently listening to is extrinsic to the CD, even though shape is a paradigmatically intrinsic property:[17]

(1) My CD is round, so *round* is either an intrinsic or an extrinsic property of my CD.

---

[15] Throughout the paper I rely on some version of the nomological deductive model of explanation, for it is the account of explanation used by Goldman in his article. There may be other ways *C* may help explain why *a* is *P* that do not involve appealing to a general connection between *C* and *P*. For example, *C* may help give a singular causal explanation of why *a* is *P* Yet nothing substantial depends on the precise nature of their connection. The argument would work just as well for any strong explanatory, explicatory, conceptual, causal, or metaphysical connection. On the limitations of the nomological deductive method, see James Woodward, "Scientific Explanation," *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta, ed., URL = <http://plato.stanford.edu/archives/fall2009/entries/scientific-explanation/>.

[16] Goldman, *op. cit.*, speaks as if this were true only of normative concepts. However, as I have cast it here, no mention or requirement of *P* being a normative property is involved.

[17] To be more precise, shape is intrinsic to most middle-sized material objects, and my CD certainly has its round shape intrinsically. See Figdor, *op cit*, Weatherson, *op cit*; Humberstone, *op. cit.*; and Brian Ellis, *The Philosophy of Nature A Guide to the New Essentialism* (Ithaca: McGill-Queen's, 2002), pp. 51–54.

(2) Assume that my CD is intrinsically round.
(3) From (2), every factor involved in accounting for my CD's being round must also be intrinsic to my CD.
(4) Presumably, the round shape of my CD can be accounted for by appealing to the fact that all points at its edge are equidistant from its center.
(5) From (3) and (4), there has to be some objective and nomological connection between the fact that the edge points of an object are equidistant from its center and the property of being round. Otherwise, the fact that all points at my CD's edge are equidistant from its center could not help explain why my CD is round.
(6) The connection between having edge points equidistant from a central point and the property of being round has explanatory power because it is an objective relation that generalizes to any round object. It is an objective truth that any object whose edge points are equidistant from its center is round.
(7) Thus, from (5) and (6), the objective and general connection between having edge points equidistant from a central point and being round cannot be intrinsic to my CD.
(8) Furthermore, from (3) and (4), this connection helps explain why my CD is round.
(9) From (7) and (8), there is at least one factor that helps explain why my CD is round that is not intrinsic to my CD.
(10) From (9), it follows that not (3); that is, not every factor that helps explain why my CD is round is intrinsic to my CD.
(11) From (10), not (2); that is, my CD is not intrinsically round.
(12) Therefore, from (1) and (11), my CD is extrinsically round.

This clearly shows that any argument that fits the schema presented in our previous section cannot be sound. In particular, it shows that Goldman's first argument against internalism in section VII of "Internalism, Externalism, and the Architecture of Justification" is unsound, since it is of this form. Thus, it also trivializes the intrinsic/extrinsic distinction upon which the internalist/externalist debate is based. *Quod erat demonstrandum.*

The purpose of this short commentary was to show that one of Goldman's arguments in "Internalism, Externalism, and the Architecture of Justification" is unsound. This goal has been achieved. However, I do not want to conclude without expressing some preliminary thoughts about the underlying mistake in Goldman's argument. In his view, determining why a subject is justified also implies determining why certain criteria explain why a subject is justified. Since, according to Goldman's account, the second of these questions cannot be answered without appealing to objective, external factors, it follows that the first question cannot be answered in an internalist way either.

Goldman recognizes that internalists may want to block this last conclusion by insisting that these later external factors should not count as genuine J-factors. However, he has also challenged the internalist to provide a principled, not *ad hoc* reason to make such a distinction. In this brief commentary, I have hinted at a reason to which an internalist may be able to appeal in order to justify her reply to Goldman: if we accept Goldman's understanding of J-rules and its concomitantly broad notion of J-factors, we could construct unsound arguments. On the other hand, if we draw the line where the internalist has insisted, we keep the intuition that all it takes for internalism to be correct is that whether or not an agent is justified in believing some proposition supervenes on the intrinsic properties of the agent.

AXEL ARTURO BARCELÓ ASPEITIA

Universidad Nacional
Autónoma de México

BOOK REVIEWS

Rescuing Justice and Equality. G. A. COHEN. Cambridge: Harvard University Press, 2008. xviii + 430 p. Cloth $50.00.

In *Rescuing Justice and Equality* (hereafter *RJE*), G. A. Cohen mounts an ambitious argument against the Rawlsian vision of a just society. The argument proceeds in two stages. Part one attempts to rescue equality from the Rawlsian willingness to countenance inequality in a just society. Part two attempts to rescue justice from the constructivist treatment of that concept, taking Rawls's constructivism as the prime example. While both parts merit careful attention, much of the material in part one has appeared in print elsewhere and will be familiar to many people. So, in what follows, I will briefly examine part one, highlighting some of the new material, and then examine part two in more detail.

Rawls's difference principle allows inequalities so long as they are needed to improve the lot of the worst off.[1] This is often taken to license giving talented people incentives to produce more, with some of the extra that they produce being used to benefit the worst off. Cohen is well known for arguing that, properly interpreted, the difference principle does not actually allow this kind of incentive-generated inequality. Part one begins by rehearsing these arguments. The basic thought is that the talented could simply choose to be more productive without receiving extra compensation and would do so if they really accepted the difference principle and applied it to their everyday lives.[2]

---

[1] This is a loose rendering of the difference principle. But this is the version Cohen largely focuses upon.

[2] This argument has been widely discussed in the literature. See, for example, Kenneth Baynes, "Ethos and Institution: On the Site of Distributive Justice," *Journal of Social Philosophy*, xxxvii, 2 (Summer 2006). 182–96; Joshua Cohen, "Taking People as They Are?" *Philosophy and Public Affairs*, xxx, 4 (October 2001). 363–86; David Estlund, "Debate: Liberalism, Equality, and Fraternity in Cohen's Critique of Rawls," *Journal of Political Philosophy*, vi, 1 (March 1998): 99–112; Thomas W. Pogge, "On the Site of Distributive Justice: Reflections on Cohen and Murphy," *Philosophy and Public Affairs*, xxix, 2 (April 2000): 137–69; Samuel Scheffler, "Is the Basic Structure Basic?" in Christine Sypnowich, ed., *The Egalitarian Conscience: Essays in Honour of G. A. Cohen* (New York: Oxford, 2006), pp. 102–29; Seana Valentine Shiffrin, "Incentives, Motives, and Talents," *Philosophy and Public Affairs*, xxxviii, 2 (Spring 2010). 111–42; Kok-Chor Tan, "Justice and Personal Pursuits," this JOURNAL, ci, 7 (July 2004): 331–62; Michael G. Titelbaum, "What Would a Rawlsian Ethos of Justice Look Like?" *Philosophy and Public Affairs*, xxxvi, 3 (Summer 2008): 289–322; Andrew Williams, "Incentives, Inequality, and Publicity," *Philosophy and Public Affairs*, xxvii, 3 (July 1998) 225–47.

0022-362X/11/0810/583–588

In *RJE*, Cohen transitions from arguing against Rawls's interpretation of the difference principle to arguing against the difference principle itself. Cohen argues that the case for the difference principle is in tension with its content. Here Cohen follows Brian Barry's reconstruction of the case for the difference principle (87). That reconstruction casts the argument in two stages. In the first stage we move from equal opportunity to equality. True equality of opportunity requires the removal of all morally arbitrary causes of inequality and "there exist no causes of inequality that are not arbitrary in the specified sense" (89). Having arrived at an equal distribution, at the second stage we see that introducing inequalities could make everyone better off. Why would anyone, especially the worst off, resist this Pareto improvement? Thus, we are lead from equality to the difference principle.

In this reconstruction, the case for the difference principle requires taking seriously an objection to morally arbitrary inequalities. But the difference principle allows inequalities that are morally arbitrary. To be sure, on Cohen's interpretation of the difference principle, this license for morally arbitrary inequalities is more circumscribed than on the standard interpretation of the difference principle. As we have seen, Cohen believes that the difference principle disallows incentives that are needed to benefit the worst off only because of the unwillingness of the talented to work more productively for equal compensation. Still, he recognizes that there might be other reasons why incentives are needed to benefit the worst off. Perhaps, for example, informational or organizational constraints prevent running the economy efficiently without incentives for the talented (155). But Cohen argues that even this more narrowly circumscribed license for morally arbitrary inequalities is in tension with the case for the difference principle: it ignores the objection to morally arbitrary inequalities that motivated the move to equality in the first place.

The role of the rejection of morally arbitrary inequalities in Rawls's view is a delicate issue and worth exploring further.[3] I will not attempt to do that here, however. I want to focus instead on how Cohen's rejection of the difference principle as a principle of justice affects the final argument of part one. So, let me turn to that argument.

As I presented Cohen's basic argument against incentive-generated inequality above, I focused on giving the talented incentives to produce more. But of course incentives are also used to encourage people to enter more socially useful occupations. One might worry, then, that Cohen's rejection of incentives can only be sustained at the cost of freedom of occupational choice. Cohen's response has a familiar form, at least at first glance: the talented could freely choose to enter more socially useful occupations. It is worth noting, though, that Cohen's rejection of the difference principle adds a wrinkle to this reply. People who endorse the difference principle might plausibly be understood to have some concern for the absolute condition of the worst off. So, perhaps such people would produce more or choose more socially useful occupations if the difference principle informed their everyday choices. But it is not clear why someone who is committed to equality rather than the difference principle should care about the absolute condition of the worst off rather than simply their relative position.[4] Hence it is not clear why someone who is committed to equality should take a more socially useful occupation.

---

[3] For further discussion of this issue, see Richard J. Arneson, "Justice Is Not Equality," *Ratio*, XXI, 4 (December 2008): 371–91; and Jon Mandle, review of *Rescuing Justice and Equality*, by G. A. Cohen, *Notre Dame Philosophical Reviews*, August 5, 2009, http://ndpr.nd.edu/news/24113-rescuing-justice-and-equality/.

Consider an example. Suppose that, given equal compensation, one prefers being a gardener to being a doctor.[5] Commitment to equality seems to rule out accepting extra compensation to take the more socially useful occupation of doctoring. But commitment to equality does not seem to settle the choice between occupations at equal compensation. Cohen sometimes seems to be sensitive to this problem, suggesting that the person who chooses the more socially useful occupation might be moved by "fellow feeling" (189, 193). But Cohen often talks as though commitment to equality alone is sufficient to motivate the choice of the socially useful occupation (182, 191, 193, 203, 204). And that seems like a mistake. Thus, once Cohen rejects the difference principle, the reconciliation of distributive justice with freedom of occupation is more complicated. Something more than an egalitarian ethos is needed.[6]

---

[4] In the context of another argument, Cohen briefly entertains the possibility that justice involves a noncomparative element in addition to the comparative element on which he focuses (322–23). Cohen treats this possibility in a skeptical tone but does not rule it out altogether. He does not consider this matter further because the argument he is focusing on in that passage does not require taking a stand on the issue. But as I indicate in the text, settling this issue is important for understanding the motivations of those who Cohen suggests ought to choose more socially useful occupations. And much of Cohen's discussion seems to run counter to taking justice to involve a noncomparative element. He even suggests that justice might, in some cases, require leveling down (316–18).

[5] When Cohen considers this case, he stipulates that doctoring does not involve any special labor burden that would justify extra compensation. It is not that the person in question hates doctoring; it is just that she prefers gardening (184).

[6] For a related discussion, see Jonathan Quong, "Justice beyond Equality," *Social Theory and Practice*, xxxvi, 2 (2010): 315–40. For a distinct objection to Cohen's treatment of freedom of occupational choice, see Michael Otsuka, "Freedom of Occupational Choice," *Ratio*, xxi, 4 (December 2008): 440–53.

The second part of *RJE* offers a diagnosis of the source of the ambivalent Rawlsian attitude toward equality that the first part of *RJE* challenges. Cohen argues that constructivism is incapable of delivering principles of justice. The kind of constructivism about justice Cohen targets holds that principles of justice gain their "normative credentials through being the product of a sound selection procedure" (274) and, in particular, that this selection procedure involves presenting specially designed selectors with the question "'What rules of governance are to be adopted for our common social life?'" (275). Cohen takes Rawls's use of the original position to be a paradigmatic instance of constructivism so understood.

There are two features of the constructivist procedure that prevent it from yielding principles of justice (275). First, the question of what rules ought to govern our common social life is properly sensitive to various non-normative facts about our situation. The deliberations of those in the original position, for example, are sensitive to facts about human nature and human society. But Cohen argues that these non-normative facts have no place in the justification of fundamental principles of justice. Second, Cohen argues that constructivism is not able properly to distinguish between justice and other values. The question of what rules ought to govern our common social life is properly sensitive to values in addition to justice, like efficiency. So, the answer to the constructivist question will yield principles that compromise justice with these other values instead of principles of justice. Cohen's discussion of both of these objections is nuanced and worthy of close examination. A quick look at the first objection, however, will suffice to highlight the depth of Cohen's disagreement with the constructivist strategy.

Cohen argues that all fundamental principles are fact-insensitive. If a fact supports a principle, Cohen contends that there is always an explanation for why this is so, and that explanation invokes or implies a more ultimate principle that would survive the denial of the fact in question. So, for example, suppose someone held the principle *we should keep our promises* on the basis of the fact that *only when promises are kept can promisees successfully pursue their projects* (234). Cohen holds that this fact could only support this principle because of some further principle, like *we should help people to pursue their projects*. In this way, examining fact-sensitive principles should lead us to uncover fact-insensitive fundamental principles. Cohen concludes that Rawls's fact-sensitive principles of justice must depend on further fundamental principles and that Rawls's constructivism obscures this.[7]

---

[7] For a close examination of the structure of this argument, see Pogge, "Cohen to the Rescue!" *Ratio*, xxi, 4 (December 2008): 454–74.

As Cohen acknowledges, this argument leaves open the possibility that the further fundamental principles in question are not themselves principles of justice (287). And he does not offer a general argument against this possibility. But it may be that the most compelling conception of the role facts play in Rawls's constructivism has this very structure. In response to Cohen's argument, for example, Arthur Ripstein has recently suggested the following interpretation of Rawls's constructivism.[8] We begin with Rawls's claim that "the correct regulative principle for anything depends on the nature of that thing."[9] This is a fact-insensitive principle and not specifically a principle of justice. From here, however, we need facts to specify the nature of the thing in question. Rawls takes principles of justice to be principles for regulating social cooperation among free and equal people. Thus, in order to identify principles of justice, we will need a specification of the problem that social cooperation poses for free and equal people. Here is where Rawls introduces the circumstances of justice, including factors like moderate scarcity and limited benevolence. These facts play a role in specifying the object that principles of justice regulate.[10] On this version of the constructivist view, then, there are no fundamental fact-insensitive principles of justice, but the justification of the principles of justice does depend on at least one fact-insensitive principle of a different kind.

This sketch of a constructivist position indicates the way in which Cohen's argument against fact-sensitive principles of justice proceeds too quickly. The argument seems to ignore the sense in which principles of justice must somehow be sensitive to the kind of beings we are in order to be principles *for us*. I think that Cohen's worry about fact-sensitive principles gains traction when he calls this into question. He asks, "Is it an axiom that human beings are *capable* of justice? Is 'original sin' a contradiction in terms?" (330). We might take Cohen to be charging the constructivist with a kind of hubris. Principles that are designed for beings like us do not leave room for a radical critique of the kind of beings we are, and it is not immediately obvious why that kind of critique should be ruled out from the start.

One might argue that "ought implies can" blocks this kind of critique. But Cohen suggests that the ultimate fact-insensitive principles have the following form: "'One ought to do *A* if it is possible to do *A*'" (251). The impossibility of doing *A* would not impeach this principle. Of course, as Cohen acknowledges, if it is impossible to do *A*, such a principle may not be of much practical interest. But Cohen claims that "the question for political philosophy is not what we should do but what we should think, even when what we should think makes no practical difference" (268).

---

[8] Arthur Ripstein, "Critical Notice of G. A. Cohen *Rescuing Justice and Equality*," *Canadian Journal of Philosophy*, xl, 4 (December 2010): 669–700. See also Samuel Freeman, "Constructivism, Facts, and Moral Justification," in Thomas Christiano and John Christman, eds., *Contemporary Debates in Political Philosophy* (Malden, MA: Wiley-Blackwell, 2009), pp. 41–60.

[9] John Rawls, *A Theory of Justice*, rev. ed. (Cambridge: Harvard, 1999), p. 25.

[10] Ripstein, *op. cit.*, p. 685.

Here I think we reach the heart of Cohen's objection to constructivism. His objection is motivated by his understanding of normative principles as fundamentally evaluative rather than action guiding. This orientation towards the normative opens up the space for questioning our capacity for justice, a question that it seems constructivism cannot entertain. While Cohen's formal arguments against constructivism may be inconclusive, he is able powerfully to press for a defense of constructivism's starting commitments, and to gesture in the direction of a different strategy.

Cohen thus leaves us with important questions about both the content of principles of justice and the concept of justice itself. While Cohen directs these questions toward Rawlsians, they are questions that anyone working in political philosophy must seriously consider.[11]

JAPA PALLIKKATHAYIL

New York University

---

[11] I am indebted to Kristi Olson, Martin O'Neill, Sam Scheffler, and James Shaw for helpful comments.

Now Available from Continuum
BY CHARLES TALIAFERRO

A *CHOICE* 'Outstanding Academic Title' of 2011

A DICTIONARY OF PHILOSOPHY OF RELIGION
Edited by Charles Taliaferro and Elsa J. Marty

A Dictionary of Philosophy of Religion is an indispensable source for students and scholars Covering historical and contemporary figures, arguments, and terms, it offers an overview of the vital themes that make philosophy of religion the growing, vigorous field that it is today. In addition to the entries co-authored by Taliaferro and Marty, leading scholars in philosophy of religion have contributed to the *Dictionary*, including Brian Davies, Pamela Sue Anderson, Paul Draper, Jerry Walls, Paul Griffiths, Douglas Hedley, Dale Jacquette, and Victoria Harrison *The Dictionary* includes a chronology, an extensive introduction to modern philosophy of religion, and a bibliography.

CHARLES TALIAFERRO is Professor of Philosophy at St Olaf College, MN.

ELSA J. MARTY is a graduate student in theology at the University of Chicago Divinity School, IL.

August 2010 | 320 Pages | PB 978 1 4411 11975 | $34.95

Part of the series *Continuum Studies in Philosophy of Religion*

THE IMAGE IN MIND
Theism, Naturalism, and the Imagination
By Charles Taliaferro and Jil Evans

The Image in Mind is a philosophical inquiry into the strengths and weaknesses of theism and naturalism in accounting for the emergence of consciousness, the visual imagination and aesthetic values Taliaferro and Evans argue that evolutionary biology alone is insufficient to account for consciousness, the visual imagination and aesthetic values. Insofar as naturalism is compelled to go beyond evolutionary biology, it does not fare as well as theism in terms of explanatory power

JIL EVANS has exhibited her paintings and prints in Minnesota, Chicago, New York City, Michigan, Memphis, Washington D.C , California, Paris, and Rome.

CHARLES TALIAFERRO is Professor of Philosophy at St Olaf College, MN.

November 2010 | 224 Pages | 6 b/w illus | HB 978 1 8470 64820 | $130.00

continuum
A Bloomsbury Company

Available from all fine bookstores
1-800-561-7704
www.continuumbooks.com

New from **ILLINOIS**

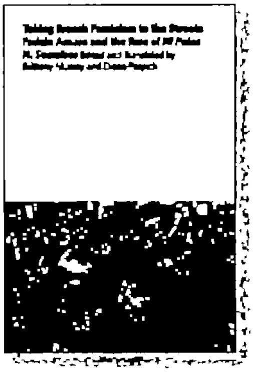
Taking French Feminism to the Streets

Fadela Amara and the Rise of Ni Putes Ni Soumises

Edited and Translated by BRITTANY MURRAY and DIANE PERPICH

With ample context and explanation of the NPNS movement for the first time for English-language readers, a discussion with the movement's founder offers additional information about immigration, discrimination, feminism, the headscarf affair, and the rise of Islamic fundamentalism.

Cloth, $65.00



The Beauvoir Series

Philosophical Writings

SIMONE DE BEAUVOIR

Edited by Margaret A. Simons with Marybeth Timmermann and Mary Beth Mader Foreword by Sylvie Le Bon de Beauvoir

"An engaging text that is sure to encourage a deeper interest in the study of Beauvoir and her place in the philosophical canon." —*Hypatia*

Cloth, $40.00

Diary of a Philosophy Student

Volume 1, 1926–27

SIMONE DE BEAUVOIR

Translation by Barbara Klaw. Edited by Barbara Klaw, Sylvie Le Bon de Beauvoir, and Margaret A. Simons, with the assistance of Marybeth Timmermann

Cloth, $40.00

"The Useless Mouths" and Other Literary Writings

SIMONE DE BEAUVOIR

Edited by Margaret A. Simons and Marybeth Timmermann. Foreword by Sylvie Le Bon de Beauvoir

Illus. Cloth, $60.00

Wartime Diary

SIMONE DE BEAUVOIR

Translation and Notes by Anne Deing Cordero. Edited by Margaret A. Simons and Sylvie Le Bon de Beauvoir. Foreword and Preface by Sylvie Le Bon de Beauvoir

"English readers are now afforded a very different portrait of the feminist philosopher approaching middle age." —*Publishers Weekly*

Cloth, $40.00

UNIVERSITY OF ILLINOIS PRESS
www.press.uillinois.edu

CAMBRIDGE

Outstanding New Scholarship in Philosophy from Cambridge University Press!

Should Race Matter?
Unusual Answers to the Usual Questions
DAVID BOONIN
$99.00 Hb. 978-0-521-76086-7
$34 99 Pb 978-0-521-14980-8
442 pp

The Evolution of Modern Metaphysics
Making Sense of Things
A W. MOORE
The Evolution of Modern Philosophy
$110 00 Hb 978-0-521-85111-4
$39.99 Pb 978-0-521-61655-3.
500 pp

CAMBRIDGE CRITICAL GUIDES

Aristotle's *Nicomachean Ethics*
A Critical Guide
JON MILLER
$85 00 Hb 978-0-521-51448-4
300 pp

Nietzsche's *On the Genealogy of Morality*
A Critical Guide
SIMON MAY
$99 00 Hb 978-0-521-51880-2
356 pp

Now in Paperback!
Leibniz
An Intellectual Biography
MARIA ROSA ANTOGNAZZA
$29 99 Pb 978-1-107-62761-1
664 pp

Prices subject to change

Ethics and the Media
An Introduction
STEPHEN J A WARD
Cambridge Applied Ethics
$90 00 Hb 978-0-521-88964-3
$28 99 Pb 978-0-521-71816-5
298 pp

Agro-Technology
A Philosophical Introduction
R. PAUL THOMPSON
Cambridge Introductions to Philosophy and Biology
$90 00 Hb 978-0-521-11797-5
$28.99. Pb 978-0-521-13375-3
258 pp

Critical Pragmatics
An Inquiry into Reference and Communication
KEPA KORTA *and* JOHN PERRY
$95.00 Hb 978-0-521-76497-1
$29.99. Pb 978-0-521-74867-4
192 pp

Proof Analysis
A Contribution to Hilbert's Last Problem
SARA NEGRI *and* JAN VON PLATO
$90 00 Hb 978-1-107-00895-3
278 pp

The Cambridge Companion to Nozick's *Anarchy, State, and Utopia*
RALF M. BADER *and* JOHN MEADOWCROFT
Cambridge Companions to Philosophy
$90 00 Hb 978-0-521-19776-2
$29 99 Pb 978-0-521-12002-9
332 pp

Human Rights Law and the Marginalized Other
WILLIAM PAUL SIMMONS
$90 00 Hb 978-1-107-01007-9
250 pp

Insurrection and Intervention
The Two Faces of Sovereignty
NED DOBOS
$85 00 Hb 978-0-521-76113-0
244 pp

Political Philosophy in the Twentieth Century
Authors and Arguments
CATHERINE ZUCKERT
$95 00 Hb 978-1-107-00622-5
$29.99 Pb 978-0-521-18506-6
290 pp

Kant, Religion, and Politics
JAMES DICENSO
$95 00 Hb 978-1-107-00933-2
302 pp

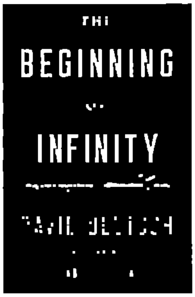
2 Volume Boxed Set
New in Paperback!
The Cambridge History of Eighteenth-Century Philosophy
KNUD HAAKONSSEN
$90 00 2 Pb books
978-0-521-28088-4 1,424 pp.

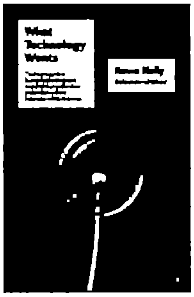
Kuhn's Evolutionary Social Epistemology
K BRAD WRAY
$95 00 Hb 978-1-107-01223-3
244 pp.

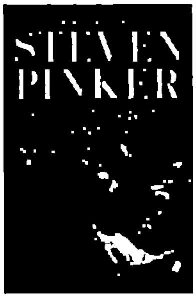
www.cambridge.org/us/philosophy

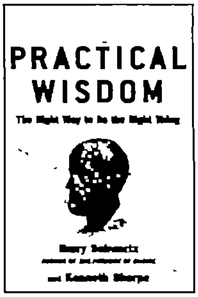
CAMBRIDGE
UNIVERSITY PRESS

NEW FROM PENGUIN GROUP (USA)









DAVID DEUTSCH
THE BEGINNING OF INFINITY
Explanations That Transform the World
A bold exploration of the nature and progress of knowledge from the author of *The Fabric of Reality*. "A philosophical exploration of progress, surprisingly lucid and thought-provoking."—*Kirkus*.
Viking • 480 pp • 978-0-670-02275-5 • $30.00

KEVIN KELLY
WHAT TECHNOLOGY WANTS
Kelly considers technology as an evolving organism with its own unconscious tendencies, following a dozen trajectories into the future to project where our relationship with technology is headed. "The best technology book I have ever read."—Nicholas Negroponte, author of *Being Digital*.
Penguin • 416 pp. • 978-0-14-312017-9 • $17 00

JIM AL-KHALILI
THE HOUSE OF WISDOM
How Arabic Science Saved Ancient Knowledge and Gave Us the Renaissance
"Lively, engaging and scholarly, al-Khalili uses his scientific background to show the importance of Arabic learning and pose challenging questions for the Muslim world today."—Hugh Kennedy, University of London.
Penguin Press • 352 pp. • 978-1-59420-279-7 • $29.95

TARIQ RAMADAN
THE QUEST FOR MEANING
Penguin UK • 224 pp. • 978-1-84-614152-2 • $24.00

STEVEN PINKER
THE BETTER ANGELS OF OUR NATURE
Why Violence Has Declined
Exploding fatalist myths about humankind's inherent violence and the curse of modernity, Pinker shows that violence has been diminishing for millennia and argues that we may be living in the most peaceful time in our existence. "A hugely exciting work and a major contribution to historiography."—Niall Ferguson.
Viking • 848 pp • 978-0-670-02295-3 • $40 00

BARRY SCHWARTZ
& KENNETH SHARPE
PRACTICAL WISDOM
The Right Way to Do the Right Thing
Two Swarthmore professors call for us to embrace what Aristotle identified as the master human virtue: practical wisdom, the ability to apply knowledge and empathy in the real world. "This book will be invaluable in a multitude of fields.... A must read."—*Choice*.
Riverhead • 336 pp • 978-1-59448-543-5 • $16 00

TONY JUDT with Timothy Snyder
THINKING THE TWENTIETH CENTURY
Penguin Press • 448 pp. • 978-159420-323-7 • $35 00
Available February 2012

PAUL JOHNSON
SOCRATES
A Man for Our Times
Viking • 240 pp. • 978-0-670-02303-5 • $25.95

PENGUIN GROUP (USA) www.penguin.com/academic
Academic Marketing Department 375 Hudson Street New York, New York 10014

THE JOURNAL OF PHILOSOPHY

VOLUME CVIII, NUMBER 11
NOVEMBER 2011

page

591 *Primitive Persistence and the Impasse between Three-Dimensionalism and Four-Dimensionalism* — Michael Della Rocca

617 *Quine Gets the Last Word* — Gary Ebbs

COMMENTS AND CRITICISM

633 *Williams on Supervaluationism and Logical Revisionism* — Nicholas K. Jones

BOOK REVIEWS

642 Solomon Feferman, Charles Parsons, and Steven G. Simpson, eds.: *Kurt Gödel: Essays for His Centennial* — Paolo Mancosu

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: David Albert, Bernard Berofsky, Akeel Bilgrami, John Collins, Arthur C. Danto, Jon Elster, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Charles Larmore, Isaac Levi, Wolfgang Mann, Frederick Neuhouser, Christopher Peacocke, Carol Rovane, Achille C. Varzi, Katja Vogt. *Consulting Editors*: James T. Higginbotham, Robert May, Charles D. Parsons, Teddy Seidenfeld, Gila Sher, Wilfried Sieg. *Managing Editor*: Alyssa Timin.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Akeel Bilgrami; *Vice President*, Michael J. Mooney; *Secretary*, Barbara Gimbel; *Treasurer*, Liaquat Ahamed; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Arthur C. Danto, Kent Greenawalt, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2012

SUBSCRIPTIONS (12 issues)

Individuals	$50.00
Libraries and Institutions	$150.00
Libraries and Institutions (print and online)	$200.00
Students, retired/unemployed philosophers	$30.00
Postage outside the U.S.	$20.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

All materials published in *The Journal of Philosophy* are copyrighted by the *Journal* and are protected by U.S. copyright law. All rights reserved. For more information about our copyright policies, please consult our January issue or contact us.

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.
POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2011 by the Journal of Philosophy, Inc ISSN 0022-362X

CALCUTTA
CENTRAL LIBRARY
CALCUTTA

THE JOURNAL OF PHILOSOPHY
VOLUME CVIII, NO. 11, NOVEMBER 2011

PRIMITIVE PERSISTENCE AND THE IMPASSE BETWEEN THREE-DIMENSIONALISM AND FOUR-DIMENSIONALISM*

An impasse. That is what we have. An impasse between so-called three-dimensionalist and so-called four-dimensionalist accounts of persistence. An impasse.[1]

Despite much vibrant work, especially of late, the leading arguments on each side of this debate seem rather ineffective. Thus, here is my regrettably breezy survey of some influential themes in the literature. The arguments for 3d'ism from the reality of change and from the notion of what it is to be the proper subject of contradictory properties:[2] the 4d'ist has perfectly coherent accounts of these matters, and the argument that these accounts are inadequate is simply question begging. The argument for 3d'ism from the claim that 4d'ism is committed to something like a counterpart theory of modality and trans-world identity:[3] this is merely a nice statement of what may be a difference between 3d'ism and 4d'ism; it is not really an argument. If the alleged commitment to counterpart theory is to provide an argument against 4d'ism, one must argue against counterpart theory itself.

---

* Earlier versions of this paper were presented at a meeting for Philosophy Faculty at Yale University, at a Philosophy Department colloquium at Harvard University, and as a Zeno lecture at the University of Leiden. I am grateful for the lively feedback on all these occasions. I am also indebted to the editors of this JOURNAL, Raul Saucedo, Sam Newlands, Laurie Paul, Jonathan Schaffer, Ruth Boeker, Justin Khoo, Jani Hakkarainen, and others for helpful comments. Finally, I owe a special thanks to the participants in my seminar at Yale in the spring of 2009 for their patience, penetration, and insightful challenges.

[1] I will hereafter use the terms "3d'ism," "3d'ist," "4d'ism," and "4d'ist" in obvious ways.

[2] See John McTaggart Ellis McTaggart, *The Nature of Existence*, ed. C. D. Broad, vol. 2 (Cambridge, UK: University Press, 1927); Sally Haslanger, "Persistence, Change, and Explanation," *Philosophical Studies*, LVI, 1 (May 1989): 1–28; and "Persistence through Time," in Michael J. Loux and Dean W. Zimmerman, eds., *The Oxford Handbook of Metaphysics* (New York: Oxford, 2003), pp. 315–54.

[3] See Peter van Inwagen, "Four-Dimensional Objects," *Noûs*, XXIV, 2 (April 1990): 245–55.

0022-362X/11/0811/591–616

And if you want to have *that* argument, I would love it. But not now.[4] Fine's argument for 3d'ism from differences between our ordinary ways of talking and thinking about parts of objects and our ordinary ways of talking and thinking about parts of events: this is a wonderfully subtle argument, sensitive to a range of linguistic and other data. But, as Fine notes, it is open to a 4d'ist to respond by "claiming that our language is deficient and our thought in error."[5] Indeed, this mass of data seems to tell us simply that the 4d'ist view may go against our intuitions in some respects. But we knew that already. What we need is a way to determine how much weight to give these pro-3d'ist intuitions and a way to measure them against any intuitions or theoretical considerations that may favor 4d'ism. The argument for 4d'ism from the "paradoxes of material constitution":[6] our intuitions about these paradoxical cases are not strong enough or consistent enough to force a decision either way on the 3d'ism/4d'ism debate or even to make us lean heavily in one direction. Sider's argument for 4d'ism from vagueness:[7] neat argument, but at most it establishes that objects have temporal parts; as we will see, a 3d'ist can easily accept that claim. Lewis's argument for 4d'ism from the problem of temporary intrinsics:[8] it is hard to see how weak or nonexistent intuitions about "just plain having" a property can decide anything of significance in metaphysics.

In the face of this welter of conflicting and differently weighted intuitions, what is most to be desired is a way to show that the intuitions that drive one theory are somehow internally incoherent or inconsistent. Anything less than this would be opposing one intuition or set of intuitions to another, and that would just continue the trench warfare that has brought us to this impasse. But if it could be shown that one theory is not simply in conflict with its rival but somehow at odds with itself, then—as long as the rival theory is not similarly internally incoherent—we would have at least the beginnings of an effective argument against the incoherent position and thus a way of breaking the impasse.

---

[4] See Michael Della Rocca, "Essentialism versus Essentialism," in Tamar Szabó Gendler and John Hawthorne, eds., *Conceivability and Possibility* (New York: Oxford, 2002), pp 223–52.

[5] Kit Fine, "In Defense of Three-Dimensionalism," this JOURNAL, CIII, 12 (December 2006). 699–714, here p. 713.

[6] Theodore Sider, *Four-Dimensionalism: An Ontology of Persistence and Time* (New York Oxford, 2003), pp 140–208.

[7] Sider, "Four-Dimensionalism," *Philosophical Review*, CVI, 2 (1997): 197–231; *Four-Dimensionalism: An Ontology of Persistence and Time*, pp. 74–139; "Temporal Parts," in Sider, Hawthorne, and Zimmerman, eds., *Contemporary Debates in Metaphysics* (Malden, MA: Blackwell, 2007), pp 241–62.

[8] David K. Lewis, *On the Plurality of Worlds* (New York: Blackwell, 1986), pp. 202–04, "Rearrangement of Particles: Reply to Lowe," *Analysis*, XLVIII, 2 (March 1988): 65–72, "Tensing the Copula," *Mind*, CXI, 441 (January 1, 2002) 1–14.

The tool that I will employ to try to break the impasse is a new characterization of the difference between 3d'ism and 4d'ism. I will bring to light a claim—namely, that persistence is primitive, inexplicable—which the 3d'ist must, by the very nature of her position, suppose and which the 4d'ist need not suppose. Then—and here is the impasse-breaking move—once we add to the mix a fundamental principle concerning branching and identity to which all, or almost all, the players on either side of the debate are deeply committed, we will be able to see that there is an internal incoherence in the position of all or almost all 3d'ists. This incoherence enables me to mount a new and, I believe, more powerful, less inconclusive argument against 3d'ism and thus for 4d'ism. However, as we will see, the incoherence that threatens 3d'ism also undermines most versions of 4d'ism. What is left at the end of these field-clearing arguments is a version of 4d'ism that rejects all cases of primitive persistence and, more generally, all cases of primitive identity. Thus the view we wind up with is 4d'ism together with the *Principle of the Identity of Indiscernibles* (PII). All other positions on persistence that accept the very popular and plausible principle about branching and identity are, I argue, internally incoherent and should be rejected.

I. A NEW CHARACTERIZATION

To begin my characterization of the debate between 3d'ists and 4d'ists, it will help to spell out ways in which they agree. First, the 3d'ist and 4d'ist agree that persistence occurs in virtue of qualitative continuity between the state at t_1 of an object, A, and the state at t_2 of an object, B. I will not here specify what this qualitative continuity amounts to. The basic idea—and this is common ground between 3d'ism and 4d'ism—is that the state at t_1 of A must not be too dissimilar from the state at t_2 of B and that any changes between t_1 and t_2 must not be too sudden. These conditions are obviously extremely vague, but the intuitive idea is clear enough, and there is no need here to fill in the account of qualitative similarity any further.

Also it is plausible to insist that when there is persistence, this is so in virtue of causal continuity between the state at t_1 of A and the state at t_2 of B. Thus if the former state is exactly like the latter state but there is no direct or indirect causal connection between these states, then, most likely, we would not say that this is a case of persistence. Here, as in the case of the similarity requirement, there is no need to make more specific the kind of causal connection required for persistence.

So the 3d'ist and 4d'ist agree that persistence occurs in virtue of causal and qualitative continuity of the state at t_1 of an object, A, and the state at t_2 of an object, B. Given this agreement, is there a substantial difference between these positions? For 3d'ism and 4d'ism to be genuinely distinct theories of persistence, they must differ as to what explains persistence. But what then can be the explanatory difference between these two theories? The crucial difference emerges from the point of agreement just outlined.

Let us say that persistence from t_1 to t_2 occurs because the state that A has at t_1 is causally and qualitatively continuous with the state that B has at t_2. For the 4d'ist, A is not identical to B; that is, persistence occurs partly in virtue of causal and qualitative continuity between the relevant states of *two* objects (A and B) that are distinct from each other. For the 4d'ist, for each period of time during which the persisting object exists there is a distinct object whose existence is limited to that time, and it is partly in virtue of this short-lived object that the persisting object in question persists.

This characterization of 4d'ism is intended to be neutral between two prominent versions of 4d'ism. The first version is *perdurantism*, according to which the persisting object is not identical to either A or B, but rather is identical to the sum of A, B, and the other relevant objects existing in the interval from t_1 to t_2. These relevant objects, in the case of perdurantism, can be seen as parts—temporal parts—of the persisting object. The above characterization is also consistent with what has been called *stage theory*. On this view, the persisting object is not the sum of A and B (and the other short-lived objects); rather, the persisting object is just A itself, and A persists in virtue of being appropriately causally and qualitatively continuous with B, which is itself a distinct persisting object.[9] My argument in this paper concerns each of the two versions of 4d'ism equally, and I will not dwell here on the difference between them.

For the 3d'ist, as for the 4d'ist, persistence from t_1 to t_2 occurs in virtue of the causal and qualitative continuity of the state at t_1 of an object, A, and the state at t_2 of an object, B. But, in contrast to the 4d'ist, for the 3d'ist the first object is identical to the second object; that is, $A = B$. Indeed, for the 3d'ist, the persisting object just is A, that is, just is *B*. Thus, *A* and *B* (that is, objects in virtue of which the persisting object persists) are not short-lived objects which are distinct from one another. Rather, the persisting object just is *A*, that is, just is *B*. For the 3d'ist, distinct short-lived objects do not explain the persistence of the persisting object. By contrast, for the 4d'ist, distinct short-lived objects do explain the persistence of the persisting object.

---

[9] Thus, because the persisting object just is the short-lived object, A, A is not—as it is according to perdurantism—a temporal part of the persisting object. Or at least it is not a proper temporal part of the persisting object. On this version of 4d'ism, these short-lived objects are called "stages," but, as with temporal parts, such stages are not happily seen as stages *of* the persisting object. This is because A is the persisting object itself. Thus if A is a stage of the persisting object, then A would be a stage of itself. Katherine Hawley, *How Things Persist* (New York: Oxford, 2001); and Sider, *Four-Dimensionalism: An Ontology of Persistence and Time*, endorse stage theory. Lewis, *On the Plurality of Worlds*; Mark Heller, *The Ontology of Physical Objects: Four-Dimensional Hunks of Matter* (New York: Cambridge, 1990); and others endorse perdurantism.

Another way to put this difference is as follows. While there may be, for the 3d'ist, an object, *A*, whose state at t_1 helps to explain persistence in this case, this object, that is, *A*, is just the persisting object itself. And while there may be, for the 3d'ist, an object, *B*, whose state at t_2 helps to explain persistence, that object, *B*, is not distinct from *A* and, like *A*, is just the persisting object itself. By contrast, for the 4d'ist, the object whose state at t_1 helps to explain persistence is not the same as the object whose state at t_2 helps to explain persistence.

To characterize the difference between 3d'ism and 4d'ism in these terms is *not* to say that, for the 3d'ist, there is no succession of short-lived objects in this case. Some 3d'ists[10] deny that in a case of persistence there are such short-lived objects, but not all do.[11] The point is that even if there are such objects for the 3d'ist, the 3d'ist denies that these objects help to explain persistence. Persistence, for the 3d'ist, does not occur in virtue of a succession of objects (even if there is such a succession). By contrast, for the 4d'ist, persistence does occur in virtue of a succession of objects. So despite the fact that the 3d'ist and the 4d'ist agree that persistence requires a series of states that exhibit causal and qualitative continuity, the 3d'ist and the 4d'ist disagree over whether a succession of distinct objects explains persistence.

Let us focus more carefully on the point of agreement, and we will see an even more illuminating way to express this fundamental difference between 3d'ism and 4d'ism. Return to the causally and qualitatively continuous states that the 3d'ist and 4d'ist both accept. For the 4d'ist, these successive states are states of *different* objects, but for the 3d'ist, these are states of the *same* object. So one way to put the difference between 3d'ism and 4d'ism is this: for the 4d'ist, the persistence of an object occurs in virtue of a causally and qualitatively continuous succession of states of *different* objects, whereas for the 3d'ist, persistence occurs in virtue of a causally and qualitatively continuous succession of states of the *same* object. Thus, for the 3d'ist, persistence of an object is explained, in part at least, in terms of a succession of states of a *single object over time*, that is, states of the persisting object. Thus Kurtz characterizes 3d'ists as understanding "change over time as a phenomenon that is nothing more than a numerically identical object instantiating different properties across time."[12] And so, for the 3d'ist, the persistence of the object is explained in part in terms of the persistence of the object itself. Thus the persistence of the object is not really explained at all; it is primitive. It is essential to the 3d'ist position that an object's persistence is thus primitive. There is no independent explanation to be had: the object persists, in part at least, because the object persists. So while, as I have stressed already, the 3d'ist says that persistence occurs in virtue of causal and qualitative continuity among states, because these states are specified as states of *the same* object over time, the 3d'ist explanation of persistence is not an independent explanation; that is, it explains persistence in terms of persistence itself. Thus the 3d'ist must see persistence as, in the end, primitive.[13]

---

[10] For example, Trenton Merricks, "On the Incompatibility of Enduring and Perduring Entities," *Mind*, xiv, 415 (July 1995) 523–31.

[11] Roxanne Marie Kurtz, "Introduction to *Persistence* What's the Problem?" in Haslanger and Kurtz, eds., *Persistence: Contemporary Readings* (New York: Cambridge, 2006), pp. 1–26, here p. 10.

By contrast, there is no such commitment to primitiveness built into or essential to 4d'ism. Thus the 4d'ist explains persistence in terms of relations of causal and qualitative continuity of *different* objects. Nothing in this account appeals to the persistence of the persisting object in order to explain the persistence of that object. Of course, in spelling out the relations of causal and qualitative continuity between the relevant short-lived objects, the 4d'ist may smuggle in the notion of the object's persistence (for example, the causal relations may be specified as holding between temporal parts of, or stages associated with, *the same object*). In that case, the 4d'ist would be, like the 3d'ist, committed to primitive persistence. But if the 4d'ist spells out the relations in terms that do not presuppose that the short-lived objects are temporal parts of, or stages associated with, a single persisting object, then he is not committed to primitive persistence. Thus the crucial difference between 4d'ism and 3d'ism over whether persistence is to be explained in terms of a succession of objects generates a further, crucial difference: the 3d'ist is essentially committed to primitive persistence; the 4d'ist is not.

Merely to point out this difference between 3d'ism and 4d'ism is not a criticism of either view. In particular, it is not a criticism of 3d'ism to argue, as I have, that 3d'ism must hold that persistence is primitive.

---

[12] *Ibid.*

[13] Cf Zimmerman, who says that for the endurantist, "no purchase is given to the argument for the existence of informative criteria of diachronic unity" Zimmerman, "Immanent Causation," in "Philosophical Perspectives 11: Mind, Causation, and World," supplement, *Noûs*, XXXI (1997): 433–71, here p. 458. See also Merricks, "There Are No Criteria of Identity over Time," *Noûs*, XXXII, 1 (March 1998): 523–31.

It is perfectly okay (as far as anything I have said so far goes) to hold, as the 3d'ist does, that persistence is primitive. My point is only that, for the 3d'ist to be consistent, she must hold that persistence is primitive. Keep that point in mind—actually, I will not let you forget it.

II. PARFIT'S PLAUSIBLE PRINCIPLE PLUS PRIMITIVE PERSISTENCE

I now want to argue—with the help of my characterization of the debate between 3d'ism and 4d'ism in terms of primitive persistence—that 3d'ism is incoherent. All I need for this purpose is a very plausible principle concerning identity and nonidentity over time.

To elicit this magical principle, I would like to turn to a famous example from Parfit. Consider a process at the beginning of which there is only one person, *A*, "in" a human body. *A*'s brain is cut in half, and each half is placed in a different and new body. At this later stage of the process, there is a person, *B*, "in" one of the bodies and a person, *C*, "in" the other body. Each of *B* and *C* is significantly and equally qualitatively (especially psychologically) and causally continuous with the original person, *A*. In light of such continuity, Parfit asks, is *A* identical to any person existing at the end of this process? In principle, there seem to be at least four possibilities.[14]

(i) $A = B$ and $A = C$.

But given that $B \neq C$ and given the transitivity of identity, (i) is not true.

(ii) *A* is identical to one of *B* or *C*, but not the other.

In this case, which one would *A* be identical to? Parfit thinks—plausibly—that each of *B* and *C* has as much claim to being *A* as the other does. After all, Parfit has stipulated that *B* and *C* are equally and significantly qualitatively and causally continuous with *A*. For this reason, Parfit denies that *one* of *B* and *C* and not the other is identical to *A*. Parfit asks rhetorically: what would make it the case that, for example, *B* and not *C* is identical to *A*? Seeing no good answer to this question, Parfit rejects this possibility. (I will return to this Parfitian rejection at length.)

(iii) *A* is identical to the sum of *B* and *C*.

This possibility is extremely problematic. *A* is a person, and it is hard to see how *B* and *C*, with their different bodies and apparent lack of coordination, can count as a (big) single person.

This leaves, for Parfit, a final option:

(iv) *A* is not identical to *B* or to *C* and, indeed, does not exist after the procedure.

---

[14] There may be others, but these four will suffice for my purposes here.

On this basis, Parfit makes the highly unusual claim that identity is not what matters in such cases. I am sympathetic to this claim, but this further point is not relevant to my attack on 3d'ism, and so that is the last we will see of it here.

It is worth noting that the same options are, in principle, available in nonpersonal cases of branching. Thus we face the same menu if *A* is an amoeba about to undergo division and *B* and *C* are the two amoebas at the end of the process.[15] In such a case, we again seem forced to conclude that *A* is not identical either to *B* or to *C*.

One can avoid this result by denying a presupposition at work in Parfit's introduction of the cases, namely, that there is only one person (or amoeba), *A*, at the beginning of the process. In denying this presupposition, one would adopt the so-called *multiple-occupancy view*. Thus there would be not one, but two people initially, *A'* and *A"*, one of whom is identical to *B* and the other to *C*. Philosophers who accept this view hold that if there is a single person (or amoeba) initially, then it cannot be identical to both or only one of *B* and *C*. But precisely because such philosophers want to avoid Parfit's conclusion that no person at the outset of the process is identical to any person at the end, multiple-occupancy theorists claim that there are more people at the beginning than we might have supposed. What is crucial here is that the proponent of this view agrees with Parfit that if there is a single person initially, then this person cannot be identical to one of *B* and *C* and not the other. It is the principle motivating this point of agreement that I want to explore.

The underlying principle is that the identity of *A* with one of *B* or *C* and not the other would be bad because there would be no ground of the identity, and thus any claim of identity would be objectionably arbitrary. Parfit expresses this worry in a number of places. First, in his original paper on this topic, he says:

> The trouble here is that...each half of my brain is exactly similar, and so, to start with, is each resulting person. So how can I survive as only one of the two people? What can make me one of them rather than the other?[16]

Later, in his book *Reasons and Persons*, Parfit frequently returns to this point. Thus, in discussing similar cases, he says:

> What could make it true that, in one case, the resulting person would be me, and in the next he would not be me? What would the difference consist in?[17]

---

[15] See Denis Robinson, "Can Amoebae Divide without Multiplying?" *Australasian Journal of Philosophy*, LXIII, 3 (September 1985): 299–319

[16] Derek Parfit, "Personal Identity," *Philosophical Review*, LXXX, 1 (January 1971): 3–27, here p. 5.

[17] Parfit, *Reasons and Persons* (New York: Oxford, 1984), p. 239

> If we are not separately existing entities, in what could this further fact consist? What could make this fact [of identity or nonidentity], in the cases in this range, either hold or fail to hold?[18]

> If personal identity does not involve a further fact, we should not believe that there are here two different possibilities: that my Replica will be me, or that he will be someone else who is merely like me. What could make these different possibilities? In what could this difference consist?[19]

> Perhaps I shall be one of the two resulting people. The objection here is that, in this case, each half of my brain is exactly similar, and so, to start with, is each resulting person. Given these facts, how can I survive as only one of the two people? What can make me one of them rather than the other?[20]

The principle at work here is that in a case of branching with significant causal and qualitative continuity, there cannot be an ungrounded identity or nonidentity between an object prior to the branching and an object after the branching.[21] More formally and more generally, the underlying principle is this:

> (PPP) In a case in which there are objects A, B, and C, $B \neq C$, B and C are equally and significantly causally and qualitatively continuous with A, and there is no other object besides A which exists at the same time as A and which is such that B and C are as causally and qualitatively continuous with it as they are with A, then it cannot be the case that $A = B$ and $A \neq C$ and it cannot be the case that $A \neq B$ and $A = C$.

I have already labeled this principle "PPP" for *Parfit's Plausible Principle.*

I name this principle for Parfit because he is so explicit about it and uses it so effectively, but it is certainly not original to him. Many other philosophers adopt or are committed to this principle. And why not? It is, after all, an extraordinarily plausible principle. Very often the principle comes up—as with Parfit's use of it—in the literature on personal identity. Thus in one of the seminal papers in the field, Williams speaks of two later branches associated with Guy Fawkes:

---

[18] *Ibid.*, p. 240.

[19] *Ibid.*, p. 242.

[20] *Ibid.*, p. 256.

[21] If (as Parfit claims) each of B and C is not identical to A, then what grounds the nonidentity of A and B, and what grounds the nonidentity of A and C? These nonidentities would be grounded in the fact that any identity of either B or C with A would be ungrounded. Notice that we cannot turn the tables and say: the identity of A and B, say, is grounded in the fact that any nonidentity of A and B would be ungrounded. Any such grounding of the identity of A and B would equally ground the identity of A and C and lead to a violation of transitivity. But the ground of the nonidentity of A and B equally applies to ground the nonidentity of A and C, and this combination does not violate transitivity. It is interesting that the claim of nonidentity here is grounded in the groundlessness of the identity. That is, nonidentity in this case is explained in terms of the inexplicability of identity in this case. Thus a crucial feature—nonidentity in this case—is explained partly in terms of explanation itself.

> We might...say that one of them was identical with Guy Fawkes, and that the other was just like him; but this would be an utterly vacuous manoeuvre, since there would be *ex hypothesi* no principle determining which description was to apply to which.[22]

In a subsequent paper, he says:

> it would be vacuous to assert one of these and abandon the other, since there is nothing to choose between them; hence the rational course is to abandon both.[23]

Williams goes on to clarify what he means by "vacuous":

> My use of the term 'vacuous' concerned not the consequences, but the grounds, of such an identification, my argument being meant to show that there would be in principle for such a case no grounds to justify a judgement of identity as against a judgement of exact similarity.[24]

Shoemaker makes a similar point about the later branches in a branching case: "it would seem arbitrary to suggest that one is identical to the original person and the other not."[25]

Nozick's rejection of a certain approach to branching shows that he too embraces PPP. First, he describes a position:

> One of Y and Z, and only one, is the continuation of the individual of which X at t_1 was a part. There is no further fact in virtue of which it is, just the luck of the draw.[26]

He then rejects this position:

> This does not seem possible, however, provided no other asymmetry is smuggled in.[27]

By denying that continuation of the individual (which, for him, is what persistence amounts to) in such a case can be a matter of

---

[22] Bernard Williams, "Personal Identity and Individuation," in *Problems of the Self* (New York: Cambridge, 1973), pp. 1–18, here p. 8.

[23] Williams, "Bodily Continuity and Personal Identity," in *op. cit.*, pp. 19–25, here p. 19.

[24] *Ibid.*, p. 20.

[25] Sydney Shoemaker and Richard Swinburne, *Personal Identity* (Oxford, UK: Basil Blackwell, 1984), p. 85. See also Shoemaker, "Persons and Their Pasts," in *Identity, Cause, and Mind: Philosophical Essays* (New York: Oxford, 2003), pp. 19–48, here p. 38.

[26] Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard, 1981), p. 62.

[27] *Ibid.*, p. 63.

"the luck of the draw," Nozick, like Williams and Shoemaker, rejects arbitrariness in such a case, denies that there can be ungrounded persistence of persons in cases of branching. Thus Nozick too embraces PPP.

Support for PPP is also strong outside the personal-identity literature. Thus van Inwagen considers the position that in the case of an amoeba undergoing fission:

> The old life is transferred to *one* of the daughter cells; the other is somehow provided with a new life.

Van Inwagen immediately and brusquely dismisses this possibility. "This is arbitrary and absurd," he says.[28]

Wiggins is no less pointed in his treatment of amoebas:

> since neither [of the resulting amoebas in a case of fission] has a better claim than the other, no sensible theory will want to count either amoeba as identical with the original one.[29]

I cannot find a clear case of a philosopher willing to deny PPP. Not even Chisholm and Swinburne—who embrace the claim that in a case of fission involving persons the earlier person is identical to only one of the two later persons—clearly reject PPP. On Swinburne's view, A is identical to B and not to C because of the presence of a soul that A, in some sense of "has," has and B has as well. Because C is without that soul, $C \neq A$.[30] Thus even on Swinburne's view, persistence and nonpersistence of persons in a case of apparent fission is grounded; in particular, it is grounded in the presence or absence of a soul. (It is such a position that Parfit refers to in one of the passages above when he speaks of the view that persons are separately existing entities.) His reliance on the soul enables Swinburne to avoid denying PPP: because of this soul, it seems natural to say that the states of A and B are causally continuous in a way that the states of A and C are not. Thus if B's psychology is due to the presence of a soul and A's psychology is due to the presence of the same soul, but C's psychology is not due to the presence of that soul, then it seems natural to say that B's states are causally continuous with A's states in a way that C's states are not. Thus B and C are not equally *causally* and qualitatively continuous with A, and so the fact that $A = B$ and $A \neq C$ does not violate PPP.

---

[28] van Inwagen, *Material Beings* (Ithaca: Cornell, 1990), p. 150.

[29] David Wiggins, *Sameness and Substance Renewed* (New York: Cambridge, 2001), pp. 72–73. See also Robinson, *op. cit.*, p. 300

[30] See Shoemaker and Swinburne, *op. cit*

Chisholm comes closest to denying PPP outright. Speaking of a person whose body undergoes fission with the result that two persons are now present, Chisholm says:

> There is no possibility whatever that *you* would be *both* the person on the right and the person on the left. Moreover, there *is* a possibility that you would be *one or the other* of those two persons.[31]

It is not clear to me, however, from this passage whether Chisholm sees the two later people as equally causally and qualitatively continuous with the earlier person. If he does not see them as equally continuous, then he is not denying PPP. If he does see them as equally continuous, then he is denying PPP.[32]

So we have not found a philosopher who clearly denies PPP. But more important than this sociological observation is the point that PPP is extremely plausible, and there seems to be exceptionally good reason to accept it not only in the case of branching amoebas, but also in the case of branching persons.

In accepting PPP, philosophers demand that there must be an explanation of certain cases of identity and nonidentity, for example, cases in which one of the two objects at the end of a process of branching is identical to the original object and the other is not. Because any trans-temporal identity in the branching case we are considering must be ungrounded and inexplicable, such identity is rejected.

Indeed, PPP seems to presuppose the more general claim that trans-temporal identity and nonidentity, as such, cannot be primitive. To see why, consider that in insisting on PPP, one claims that it cannot be that $A = B$ (say) and $A \neq C$ and that this is so *because* such identity and nonidentity would be arbitrary, primitive. In affirming PPP, we reject certain claims of identity and nonidentity precisely because the identity and nonidentity would be primitive. And thus in affirming PPP, we seem to identify primitive identity and nonidentity as the culprit, as something that is wrong and to be avoided.

That a general rejection of primitive persistence is at work in PPP also emerges in this way. If, in a case in which A and B exist at different times, $A \neq B$ in part because an identity here would be primitive, then it follows that in cases where there is an identity of two objects,

---

[31] Roderick Chisholm, *Person and Object A Metaphysical Study* (Chicago: Open Court, 1976), p. 111. See also Chisholm, "Reply to Strawson's Comments," in Howard E. Kiefer and Milton Karl Munitz, eds., *Language, Belief, and Metaphysics* (Albany: SUNY, 1970), pp. 187–89.

[32] See Shoemaker, "Persons and Their Pasts," p. 38n19, for similar uncertainty about Chisholm's view Merricks, "There Are No Criteria of Identity over Time," raises the possibility that something like PPP is false, but he does not deny PPP outright.

x and y, that exist at different times, that trans-temporal identity holds in part precisely because the identity of x and y is not primitive. In other words, if a case of trans-temporal identity (between A and B) is rejected simply because it would be primitive, then any case of trans-temporal identity holds in part because that identity is not primitive. And if trans-temporal identity in general must not be primitive, then so too trans-temporal nonidentity in general must not be primitive.[33] So PPP seems to presuppose a general commitment to the rejection of primitive trans-temporal identity and nonidentity.

You might try to avoid this result by saying that whereas primitive identity and nonidentity in branching cases is bad, in standard cases (without branching) primitive identity and nonidentity is okay. But this move seems ad hoc and ineffective. If identity in the branching cases is bad (as PPP insists), then why is it bad? It seems to be bad simply because the identity and concomitant nonidentity would be primitive, because there would be nothing in virtue of which the identity and nonidentity would hold. But if primitive identity and nonidentity are not to be tolerated in cases of branching, then why should they be tolerated in cases where there is no branching?

The best way, perhaps, to try to avoid the rejection of primitive trans-temporal identity in general is by drawing the following distinction between persistence in a nonbranching case and persistence in a branching case. One might argue that claims of primitive trans-temporal identity and nonidentity in branching cases would be *arbitrary* in a way that claims of primitive persistence in standard, nonbranching cases are not. In the branching cases, if one affirms that A and B are identical (and denies that A and C are identical), that would seem completely arbitrary, for it seems that we could equally have said that $A \neq B$ and that $A = C$ instead. (Williams, Shoemaker, and van Inwagen, as we have seen, all appeal to the notion of arbitrariness in this context.)

---

[33] To show that if trans-temporal identity must not be primitive, then trans-temporal nonidentity must also not be primitive, assume the opposite. That is, assume that trans-temporal identity must not be primitive, but trans-temporal nonidentity may be primitive If this is the case, then if x (which exists at t_1) is identical to y (which exists at t_2), then there must be a qualitative ground of that identity. In other words, sameness of qualitative properties guarantees identity over time. But if, for each identity over time, there must be a qualitative ground, then there must be a qualitative ground for each nonidentity over time too. If x and y are not identical, this nonidentity must be explained not by the nonidentity of x and y, but by the fact that there is a qualitative difference between x and y If there were no qualitative difference between x and y, then, given that identity is not primitive, x and y would be identical. Thus, if x and y are not identical, then there must be a qualitative difference between them after all But this contradicts our initial assumption in this note, and so we must hold instead that if trans-temporal identity must be grounded, then so too is trans-temporal nonidentity That is, if trans-temporal nonidentity is primitive, then so too is trans-temporal identity

However, in cases of ordinary, nonbranching persistence with considerable causal and qualitative continuity, there is no such arbitrariness. We could not just as easily have said that an earlier object, *D*, is not identical to a later object, *E*, for what alternative to *E*'s being identical to *D* could we have proposed in a nonbranching case? (Imagine that *D* is the chair in the room at t_1, *E* is the chair in the room at t_2, and there is no funny business going on.)

But this response will not work. This is because there *is* an equally good alternative to affirming *E*'s identity with *D*; there *is* arbitrariness in such a case, as long as one accepts that standard, nonbranching cases of persistence may involve primitive trans-temporal identity. If primitive trans-temporal identity is allowed, then one must also allow primitive trans-temporal nonidentity.[34] Given that primitive identity and nonidentity are allowed, it could be that, no matter how much causal and qualitative continuity *E* exhibits with *D*, *E* is just another object distinct from *D*, and primitively so: *E* is distinct from *D* simply because *E* is distinct from *D*, in the same way that if *D* is primitively identical to *E*, this identity holds simply because $D = E$. Given the legitimacy of primitive identity and nonidentity in nonbranching cases, it seems arbitrary to suggest that the later object is identical to the earlier one instead of saying that the later object is distinct from the earlier one. What reason could one have for preferring the identity of *D* and *E* to their nonidentity, unless one thought that nonbranching causal and qualitative continuity by itself was enough to guarantee identity? But if one thinks this, then the identity of *D* and *E* would not be primitive after all. Indeed, it would be grounded in the nonbranching causal and qualitative continuity between *D* and *E*. Thus if standard, nonbranching persistence may be primitive, then there is arbitrariness in standard, nonbranching cases. And so, if standard, nonbranching persistence may be primitive, then the purported distinction between the branching and nonbranching cases—namely, that affirming identity in a branching case is arbitrary in a way that affirming identity in a standard, nonbranching case is not—vanishes. But remember that the proponent of PPP who wants to avoid a blanket rejection of primitive persistence needs to find a relevant distinction between the branching and nonbranching cases. Given the point that if standard, nonbranching persistence is primitive, then there seems to be no relevant distinction between branching and nonbranching persistence (each case would bring with it arbitrariness), it follows that the proponent of PPP who wants to avoid a blanket rejection of primitive persistence must deny that ordinary persistence may be primitive. But if one does this, then as long as one also denies that there can be primitive identity in the branching cases (that is, as long as one accepts PPP), one is committed to denying primitive persistence in both branching and nonbranching cases. Thus one is committed to a blanket rejection of primitive persistence after all. So the point stands: given PPP, one is committed to denying primitive persistence in general.

---

[34] Assume that trans-temporal identity is primitive, but trans-temporal nonidentity is not primitive. If this is the case, then if x (which exists at t_1) is not identical to y (which exists at t_2), then there must be a qualitative ground of that nonidentity. But if for each nonidentity there is a qualitative ground, then there must be a qualitative ground for each identity after all, namely, the lack of a qualitative ground of nonidentity. Thus if $x = y$ this identity must be explained not by the identity of x and y, but by the fact that there is no qualitative difference between x and y. Thus there is a qualitative ground of identity. But this contradicts our initial assumption, and so we must hold instead that if trans-temporal nonidentity is grounded, then so too is trans-temporal identity. That is, if trans-temporal identity is primitive, then so too is trans-temporal nonidentity.

Of course, on seeing that PPP leads to a blanket rejection of primitive trans-temporal identity and nonidentity, you may choose to give up PPP. But this seems to be an unhappy path to take. As I have stressed, with the backing of many of the bigshots of the personal identity and persistence literature, PPP is extremely plausible. (It is not called P*P*P for nothing.) If you take this unhappy route, I have nothing further to say to you. I am not primarily directing my argument here to one who may reject PPP; rather, I am trying to elicit what proponents of PPP (that is, just about all philosophers, as far as I can see) are committed to. And what I have been arguing is that proponents of PPP (that is, just about all of us) are committed to the rejection of primitive trans-temporal identity and nonidentity in general.

III. 3D'ISM UNDERMINED

PPP and the resultant commitment to the rejection of primitive persistence in general leads to big trouble if one tries to combine PPP with 3d'ism. As we have seen, 3d'ism by its very nature embodies a commitment to primitive trans-temporal identity and nonidentity. Given this commitment and given the commitment to rejecting primitive trans-temporal identity and nonidentity at work in PPP, the 3d'ist cannot coherently accept PPP. Any 3d'ist who accepts PPP has an incoherent overall position. This is unfortunate, because it seems all or almost all philosophers, including all or almost all 3d'ists, accept PPP. But, again, the key point is not about the prevalence of the acceptance of PPP. The key point is that PPP is extremely plausible, and as we can now see it puts us in possession of a new and powerful argument for 4d'ism.

Simply put, the argument is: PPP entails the denial of 3d'ism. Less simply put, the argument is: given PPP and its commitment to the rejection of primitive trans-temporal identity and nonidentity, and given that 3d'ism is committed to the legitimacy of primitive trans-temporal identity and nonidentity, it follows that 3d'ism must be false. (It also follows that, if there is persistence at all, then 4d'ism is true.)

I am certain that 3d'ists and even many 4d'ists are champing at the bit to argue against me. So let me make it easy for you: all you need to do is to offer good reasons for rejecting *one* of the following three claims:

(a) PPP;
(b) PPP commits one to the rejection of primitive trans-temporal identity and nonidentity in general;
(c) 3d'ism commits one to the legitimacy of primitive trans-temporal identity and nonidentity.

I have already argued, in section II, for (b) and, in section I, for (c). Further (a), that is, PPP, is as I have stressed difficult to challenge. Thus, I believe, there are no good options open to the 3d'ist for defending her position.

With 3d'ism thus behind us, the question becomes: which version of a non-3d'ist account of persistence should we accept? This question is salient in light of the fact that, as we will see, the same PPP-based reasoning at work in the rejection of 3d'ism also undermines at least some versions of 4d'ism.

IV. SPINNING DISKS AND 4D'ISM

To see why, let us turn to a case that has figured prominently in the debate over persistence, usually in attempts to support 3d'ism. For reasons that will soon emerge, I think that this example does *not* support 3d'ism, but I want to employ this case not in order to adjudicate the debate between 3d'ism and 4d'ism. Rather, my aim is to reveal which alternative to 3d'ism is most coherent by discerning the limited options available for addressing the example.

Here is the case, pared to its essentials.[35] Imagine a universe that contains nothing but two eternal, perfectly homogeneous, perfectly circular disks which are made of exactly the same kind of matter and which have exactly the same dimensions. In this universe, there never has been and never will be anything else besides these two disks. Such a (boring) universe seems possible. At least, let us assume for now that it is possible. I will return to this assumption later. Let us make the further assumption that one of the disks, *A*, is spinning at t_1, and the other, *B*, is not. Given this assumption, the question of the moment is: what can ground the spinning of *A* and the being-at-rest of *B*? Keep in mind that the disks, as homogeneous, have no further difference between them in terms of which to ground the difference with regard to motion. There is no deformity or discoloration or distortion in one disk at t_1 that can differentiate it from the other disk and serve to ground the difference with regard to motion. Nor is there any difference elsewhere in the universe to which we might appeal to give an account (in terms extrinsic to the disks) of the motion or rest of the disks. Sider, in other cases, appeals to extrinsic differences to ground kinematic facts about the disks, but he acknowledges that in a world with nothing else besides the disks (or in an otherwise completely symmetrical universe), there would be no such grounding for facts about motion.[36] Nor can we appeal to any feature at t_1 such as instantaneous velocity or vector quantity to ground the state of motion of the disks. For what would such features be grounded in—in virtue of what would one disk have a different vector quantity or different instantaneous velocity from the other?[37]

---

[35] The case stems from Kripke's unpublished lectures on identity and from David Armstrong, "Identity through Time," in van Inwagen, ed., *Time and Cause. Essays Presented to Richard Taylor* (Boston: Reidel, 1980), pp. 67–78.

In this universe, what, if anything, grounds the state of motion of the disks? No qualitative difference between the disks seems to be available to do this job. Perhaps all that we can appeal to here are identity facts. The question of whether a disk is spinning or at rest is really the question: is the right side (say) at t_1 identical to the right side at t_2? If not, then the disk is spinning (or at least the disk is moving somehow) at some point during this period. If the sides are identical, then the disk is not spinning.[38] Lewis states this connection between spinning and identity over time, "Whether the disk spins is, of course, definable in terms of the persistence of its parts through time."[39]

But even if the spinning could be grounded in identity facts in this way, these facts themselves would have to be ungrounded or primitive. Thus, take the case of the spinning disk *A*: the right side at t_1 is not identical to the right side at t_2. In virtue of what does this lack of persistence hold? Is there any qualitative difference between the right side of *A* at t_1 and the right side of *A* at t_2 in virtue of which to ground the nonidentity of the sides? Given the homogeneity of the disk, it seems that there is not. That there is no such ground is especially evident when we compare *A* with *B*, the stationary disk. The right side of *B* at t_1 is, of course, identical to the right side of *B* at t_2, but given the homogeneity, the corresponding sides of *A* at those two moments are indiscernible from the sides of *B* at those two moments. Any feature that would purportedly ground the persistence of *A*'s sides would be shared by *B*, the stationary disk. So the persistence of *A*'s sides could not be grounded after all. There is no ground for facts about persistence and the failure of persistence in this case. If we hold that there is a fact of the matter in such a case as to whether the disk is spinning, then the persistence of the sides of the disks must be primitive. Lewis realizes this point when he speaks of the spinning disk immediately after the passage just quoted: "so in the first instance it is persistence that fails to supervene."[40]

---

[36] Sider, *Four-Dimensionalism: An Ontology of Persistence and Time*, pp. 230–36.

[37] *Ibid.* and Zimmerman, "One Really Big Liquid Sphere: Reply to Lewis," *Australasian Journal of Philosophy*, LXXVII, 2 (1999): 213–15 are helpful in exploring why such features do not enable us to ground the motion of the disks.

[38] More precisely, if the right side of the disk at t_1 is identical to the right side at t_2, and if, for any two moments between t_1 and t_2, the right side of the disk at one of those two moments is identical to the right side at the other and is identical to the right side at t_1, then the disk is not spinning.

[39] Lewis, "Introduction," in *Philosophical Papers*, vol. 2 (New York: Oxford, 1986), pp. ix–xvii, here p. xiii note 5.

Thus if one accepts the possibility that there can be homogeneous disks (in a symmetrical or otherwise empty universe) and if one accepts the further possibility that there can be a fact of the matter as to whether such a disk is spinning or at rest, then one is committed to primitive persistence in this case. Equivalently, if one rejects the possibility of primitive persistence in this case, then one either must deny that there is a fact of the matter as to whether a disk in such a situation spins or is at rest or reject the possibility that there can be homogeneous disks (in a symmetrical or otherwise empty universe).[41]

In light of this connection between the disk example and primitive persistence, we can lay out the possible responses to this example as follows:

(1) 3d'ism with primitive persistence in the universe of homogeneous disks.

As we have seen, the 3d'ist is committed to primitive persistence, so it is a bit redundant to say "3d'ism with primitive persistence" as if "with primitive persistence" adds more information. But the redundancy serves to illustrate why 3d'ists tend to have no problem with the disk example: since, as we have just seen, the spinning in the disk example turns on the possibility of primitive persistence, in asking the 3d'ist to accept the disk example one is not asking her to accept anything she is not already committed to.

(2) 4d'ism with primitive persistence in the universe of homogeneous disks.

As we have seen, 4d'ism is compatible with primitive persistence, so this response to the example seems to be coherent.

(3) 4d'ism with no primitive persistence in the universe of homogeneous disks.

---

[40] *Ibid.*

[41] Sider is explicit on this point. See *Four-Dimensionalism: An Ontology of Persistence and Time*, pp. 233–36

As we have also seen, 4d'ism is compatible with the rejection of primitive persistence. This third combination of views is committed to the claim that, in such a universe, there is no fact of the matter as to whether a disk is spinning.

Notice that a fourth option, namely, 3d'ism with no primitive persistence, is not really an option, because again the very nature of 3d'ism requires that persistence is primitive. Thus the three options above are exhaustive, or at least they exhaust the ways 3d'ists and 4d'ists can respond to this example, on the assumption that there can be a universe of homogeneous disks.

In light of these options, we can categorize various philosophers with regard to this example. All 3d'ists should, of course, accept option (1). Sider clearly commits himself to option (3), 4d'ism with no primitive persistence in this case. He thus accepts that there is no fact of the matter as to spinning in this case.[42] Sider's position here is more consistent than Lewis's. Lewis, of course, accepts 4d'ism and Humean supervenience—the view that all truths are grounded in qualitative variation among space-time points.[43] Lewis also holds that there is a legitimate Humean ground of the facts concerning spinning in this case.[44] Thus, for Lewis, the persistence in this case is not primitive; (3) is true, and there is a fact of the matter as to spinning in this case. However, as I have argued and as Sider and Zimmerman show in detail, there does not, contra Lewis, seem to be a legitimate Humean ground of spinning and not spinning in this case.[45]

Hawley can be seen as accepting (2). She embraces 4d'ism, but, unlike Sider, she holds that there can be a fact of the matter as to spinning in such a case. She holds that the fact of spinning consists in what she calls *nonsupervenient relations* between the temporal parts of the disk (or stages associated with the disk), relations that do not supervene on, and hence are not explained by, intrinsic features of these temporal parts or stages. I believe that these nonsupervenient relations commit Hawley to the primitiveness of spinning and so to primitive persistence in this case.[46]

Armstrong's position is a bit more difficult to characterize. On the one hand, he is a 4d'ist who accepts the possibility of such disks and that there can be a fact of the matter about their spinning. This seems to commit him to denying (3), which requires that there be no fact of the matter in this case. But Armstrong does see persistence and spinning as grounded in causal facts. So it seems that Armstrong does not accept primitive persistence in this case; instead, for him, persistence is grounded in causal facts. This might seem to put Armstrong among proponents of (3). But I think this would not be a proper way to categorize him. This is because Armstrong sees causation as a primitive relation, and so he might see the primitiveness of causation as inherited by persistence. Indeed, Lewis suggests as much in the footnote that I have already invoked; in describing a position (not his own) in which the persistence of parts of the disk fails to supervenene, he says, "But that might be because causation fails to supervene, and persistence requires causal continuity."[47] In any event, it is not important to see Armstrong (or Lewis or any other 4d'ist) as taking option (2) or as taking option (3), because I will now argue that both options are incoherent.

---

[42] *Ibid.*, pp. 233–34

[43] Lewis, "Introduction," in *Philosophical Papers*; and "Humean Supervenience Debugged," *Mind*, CIII, 412 (October 1994) 473–90.

[44] See Lewis, "Zimmerman and the Spinning Sphere," *Australasian Journal of Philosophy*, LXXVII, 2 (1999): 209–12.

[45] See Sider, *Four-Dimensionalism: An Ontology of Persistence and Time*, pp. 209–36; Zimmerman, "One Really Big Liquid Sphere: Reply to Lewis"

[46] See Hawley, "Persistence and Non-Supervenient Relations," *Mind*, CVIII, 429 (January 1999): 53–67; *How Things Persist*.

First note that, given PPP, option (1) is ruled out, and so is that version of 4d'ism that is option (2). Both options (1) and (2) involve primitive persistence, which is incompatible with PPP. What then is left? Option (3): 4d'ism with no primitive persistence of the homogeneous disks. On this option, as we saw, if the disks exist, there can be no fact of the matter as to whether the disk is spinning or at rest.

But is this the route that the 4d'ist has to take, given PPP and the rejection of primitive persistence? Perhaps not. There may be another option, and indeed it is forced on us by the kind of reasoning at work behind PPP. This reasoning undermines a presupposition shared by each of the three responses to the spinning-disk case, namely, that it is possible for there to be perfectly homogeneous disks in an otherwise empty or symmetrical universe. To see why such a presupposition gets knocked out, let us look more closely at the rejection of primitive trans-temporal identity and nonidentity.

To demand that persistence not be primitive is to insist on identity-free explanations of persistence; that is, it is to insist that in a case of persistence, *A* persists in virtue of some fact beyond the very persistence of *A* itself. By contrast, any view according to which persistence is primitive is one that explains persistence in a way that is *not* identity free.

The insistence on identity-free explanations of persistence is similar to the demand for identity-free explanations of trans-world identity.

---

[47] Lewis, "Introduction," in *Philosophical Papers*, p. xiii note 5.

To insist on an identity-free explanation here is to insist that identity across worlds not be primitive, that it be grounded in some fact beyond the fact of trans-world identity itself. One way to do this (perhaps there are others) is to ground trans-world identity in terms of trans-world similarity (whatever that is). We see this approach in certain counterpart-theoretic accounts of trans-world identity. The fact that A in w_1 is identical to B in w_2 consists in the fact that A and B have the requisite degree of similarity.[48] It is not surprising that Lewis and Sider—both proponents of counterpart-theoretic accounts of modality and thus of identity-free accounts of trans-world identity—also espouse identity-free accounts of trans-temporal identity. In each of these cases of identity, Lewis and Sider are committed to the identity in question not being primitive. (We will see later, however, that Lewis and Sider may not be consistent in carrying out this commitment.)

In contrast to an identity-free account of trans-world identity, there are broadly Kripkean approaches to this topic. On this kind of account, trans-world identity is not, and cannot be, accounted for in identity-free terms. For Kripke, there is no point in appealing to similarity in order to find a ground for identity across worlds, a ground beyond the simple fact that an object, x, is identical to x across worlds. And it is not surprising that those committed to a Kripkean account of trans-world identity (such as, say, Kripke) tend to be committed to 3d'ist accounts of persistence. Both views are wedded to the legitimacy of explanations of identity and nonidentity that are not identity free, that are primitive. Thus these positions are, in this respect, of a piece. However, in saying this, I am not saying that if one is a 3d'ist one must accept a Kripkean account of trans-world identity or vice versa. Nor, on the flip side, am I saying that there is any direct entailment between 4d'ism and counterpart-theoretic accounts of trans-world identity. These matters are, of course, very complicated.[49] I am merely pointing out that the commitment to identity-free explanations of identity in the trans-temporal case puts pressure on one to demand identity-free explanations of identity in the trans-world case.

In light of the demand by a proponent of PPP for an identity-free explanation of trans-temporal identity and nonidentity (and in light of the demand, made by many, for identity-free explanations of trans-world identity), it is natural to expect that identity *at a time* also must be explained in an identity-free, nonprimitive way. If we do not want the nonidentity of *A* and *B* to be primitive when *A* and *B* exist at different times, if we want to be able to say more than that $A \neq B$ because $A \neq B$, then similarly we would want to be able to say more than that $D \neq E$ (in a case in which *D* and *E* exist at the same time) because $D \neq E$ and that this is all that one can say by way of explaining that *D* and *E* are not identical..

---

[48] Of course, there are lots of qualifications here concerning constancy and inconstancy of modal predication that I cannot talk about here. See, again, Della Rocca, "Essentialism versus Essentialism."

[49] See Hawthorne, "Three-Dimensionalism," in *Metaphysical Essays* (New York: Oxford, 2006), pp. 85–109

So it seems that one who accepts PPP and thus insists on the illegitimacy of primitive persistence is under pressure to hold that primitive identity and nonidentity at a time are also illegitimate. As before, I am not saying definitively that one cannot coherently hold that persistence is not primitive, that the explanation of persistence must be identity free, and at the same time countenance primitive identity and nonidentity at a time, that is, countenance explanations of identity at a time that are not identity free. Perhaps one can coherently combine these views, but I do not see how this can be done. More importantly, there is considerable pressure on one who rejects primitive persistence to reject primitive identity at a time as well. At least the burden is on the one who accepts primitivity in one of these cases and not in the other to offer a good reason that this discrepancy should be the case. That is, such a person would need to give a principled reason for treating identity at a time as primitive while holding that identity over time (or across worlds) is not primitive.

Let us see what happens if we—in light of PPP and the concomitant rejection of primitive persistence—also reject primitive identity and nonidentity at a time. Forget the issue of whether the disks are spinning, and focus instead on a point agreed on by all three responses to the spinning-disk case, namely, that there could be a homogeneous disk all by itself in the universe. Let us focus on a specific time, t_1, and consider the left half of the disk at this time. Call the left half "Larry." Focus also on the right half, Ralph. Obviously, the left half of the disk at t_1 is not identical to the right half at t_1. This is a case of nonidentity at a time. If we are to reject primitive identity and nonidentity at a time, then we must be able to give an explanation of this nonidentity. But in virtue of what is Larry $\neq$ Ralph? They are *ex hypothesi* qualitatively exactly alike. They are, to use the misleading lingo, indiscernible. They are intrinsically exactly alike (whatever "intrinsic" means), and they have all the same extrinsic properties—each is connected to a half-disk with such-and-such qualities, and so on. The only way to explain this nonidentity at a time is to say that Larry is not Ralph because Larry is not Ralph. There is no further difference to appeal to. But to explain this nonidentity in this way is to treat it as primitive and so is not really to explain it at all. However, in rejecting primitive identity and nonidentity at a time, we were rejecting such nonexplanations of identity and nonidentity. And so the possibility of homogeneous disks in a symmetrical or otherwise empty universe must be rejected too.

"But not so fast," I can hear you saying. "We *can* explain the nonidentity of Larry and Ralph in an identity-free way. After all, Larry and Ralph occupy different spatial locations, and we can illuminatingly individuate Larry and Ralph in terms of these different locations."

Okay, but this only pushes the problem one step back, for now we can ask: in virtue of what are these locations not identical? (This would be a case of nonidentity at a time.) Surely, these locations are qualitatively exactly alike, and we cannot appeal to the different objects at the locations in order to individuate the locations because we just got through trying to individuate the objects by appealing to the locations. We cannot noncircularly carry out both acts of individuation. We then seem to be forced to say that one location is different from the other simply because one is different from the other.[50] This would be individuation of locations that is not identity free, and, again, such individuation would be precluded by the rejection of primitive identity and nonidentity at a time.

Thus if—in keeping with the lessons of PPP and the rejection of primitive persistence—we also reject primitive identity and nonidentity at a time, then we must deny that Larry really is not identical to Ralph. That is, Larry must be identical to Ralph after all. This collapse of apparently different halves can continue indefinitely. The top half of the remaining half is indiscernible from, and hence identical to, the bottom half of that half, and so on. Ultimately, we would be left with a single point and not a disk or half a disk at all. Thus once we accept PPP and draw out its implications, the presupposition of the spinning-disk example—that there can be homogeneous disks in a symmetrical or otherwise empty universe—must be given up.

So even the third option for responding to the spinning-disk case must be rejected. But how can we do this? Option (3) consists of three claims:

(i) A homogeneous disk in an otherwise empty universe is possible.
(ii) 4d'ism is true.
(iii) There is no primitive persistence in the case of the homogeneous disk.

As we have seen, PPP entails 4d'ism (as long as there is persistence at all), and PPP also entails that there is no primitive persistence. So, given PPP, the offending party must be (i), the possibility of a homogeneous disk or, more generally, the possibility of objects in symmetrical universes. Given PPP, we must reject option (3), and given PPP, if we reject option (3), we must reject (i); that is, we must reject the claim that a homogeneous disk in an otherwise empty universe is possible.

---

[50] For more on this kind of maneuver, see Della Rocca, "Two Spheres, Twenty Spheres, and the Identity of Indiscernibles," *Pacific Philosophical Quarterly*, LXXXVI, 4 (December 2005): 480–92.

That this fate should befall the view to which Sider and Lewis are committed (namely, 4d'ism with no primitive persistence in the case of the homogeneous disks) is ironic, because that view is motivated by the desire to find nonprimitive, identity-free explanations of trans-temporal identity. But Lewis and Sider run afoul of their own commitments here, because the commitment to no-primitive-persistence commits them to the rejection of primitive identity and nonidentity at a time. In accepting the possibility of homogeneous disks, Lewis and Sider accept—contrary to their own commitments—a case of primitive identity and nonidentity at a time. In general, the incoherence of Lewis and Sider's position is to insist on identity-free individuation across times while not realizing that this insistence on identity-free trans-temporal individuation (an insistence embodied in PPP) leads to the requirement that the explanation of identity and nonidentity in general be identity free, nonprimitive.

V. 4D'ISM. PII.

So where are we? We have seen the cascading elimination of each of the three options I outlined earlier for dealing with the homogeneous disks. Options (1) and (2) are eliminated because they conflict with the commitment to identity-free explanations of persistence, a commitment brought on by the acceptance of PPP. (Reminder to my opponents: you can make me stop just by providing good reason to deny PPP. But go ahead. Try to deny it. Make my day.) Option (3) is rejected because, while it admirably insists on an identity-free explanation of persistence in the case of the disks, it nonetheless and incoherently accepts the possibility of distinct objects in a symmetrical universe, objects that cannot be individuated in an identity-free way.

In order to progress, in addition to giving up primitive persistence, we need to give up a presupposition of each of the three responses, namely, the presupposition that there can be symmetrical universes, that there can be individuation of objects at a time that is not identity free. If we insist that identity is thus in general nonprimitive, then we get rid of the 3d'ist's primitive persistence, those versions of 4d'ism that allow for primitive persistence, and finally those versions of 4d'ism that—while sympathetic to the view that identity in general should not be primitive—nonetheless and incoherently allow for primitive nonidentity in certain cases.

What then is left? Answer: the general requirement that identity and nonidentity not be primitive, and thus the acceptance of a version of 4d'ism combined with PII, the *Principle of the Identity of Indiscernibles*.

Is this position also incoherent or otherwise unacceptable? It is hard to see why it should be. I grant that this position is implausible: who would find plausible the view that there cannot be symmetrical universes? Who would find plausible the view that persistence occurs in virtue of distinct short-lived objects that must not be indiscernible? But we have been led to this view by a series of plausible steps, beginning with PPP. Further, if the most important requirement for a philosophical theory is to avoid incoherence, then this view—4d'ism with PII—seems to have the advantage over all of its competitors that we have examined.

Nevertheless, I cannot rule out the possibility that this position too may harbor some internal incoherence. For a presupposition of this position is that there can be a multiplicity of distinct objects (in particular, the multiplicity of short-lived objects in virtue of which persistence occurs). And it may well be that the same demand for explanation that leads to the rejection of primitive persistence and of primitive identity and nonidentity generally also leads to the rejection of a multiplicity of objects. But this is a topic for another occasion, on which I will subject the version of 4d'ism at which I have arrived to deeper analysis.

Yet even apart from a full examination of this new version of 4d'ism, we have achieved some important results:

(1) We have shown how 3d'ism is—given a plausible further principle (PPP)—incoherent, and we thus have devised a powerful and new argument for 4d'ism.
(2) We have, on the same basis, rejected all versions of 4d'ism, such as perhaps Hawley's, that allow for primitive persistence.
(3) We have rejected Lewis and Sider's versions of 4d'ism, which—while sympathetic to the denial of primitive persistence—nevertheless incoherently allow for primitive identity and nonidentity.
(4) Faced with these dwindling options, we have been led—surprisingly and for previously unexplored reasons—toward PII, the view that there is no difference between objects without a qualitative difference, that no objects are individuated primitively, and thus toward a combination of 4d'ism with PII.

Each of these four results is made possible by a further and more fundamental result:

(5) We have offered a new characterization of theories of persistence in terms of their acceptance of primitive persistence.

This characterization has opened up a new way of evaluating theories of persistence and of identity more generally. In examining such theories we must always keep in mind the question: what commitments do the theories in question have to primitive identity or to the denial of primitive identity, and how consistently do these theories carry out these commitments? It is by relentlessly wielding this tool provided by my characterization of the debate that we have rejected the three most common responses to the spinning-disk case and undermined 3d'ism, with the result that we are left—for the moment at least—with a surprising combination: 4d'ism. PII.

MICHAEL DELLA ROCCA

Yale University

QUINE GETS THE LAST WORD*

Today, fifty years after the first publication of W. V. Quine's *Word and Object*,[1] two related criticisms of Quine's philosophy have become entrenched. One is that Quine's naturalistic epistemology is descriptive, not normative, and so fails to provide an account of epistemic justification. The other is that Quine's naturalistic account of meaning leads to the absurd conclusion that our sentences are meaningless, so we cannot use them to make assertions. I shall argue that these entrenched criticisms ultimately rest on a failure to grasp Quine's resolutely minimalist understanding of language use.

In Quine's view, language use—in particular, our use of language to theorize—is pragmatically indispensible, and although our sentences are associated with sensory stimulation by the mechanism of conditioned response, descriptions of such associations do not convey what it is to use one's sentences to make assertions. When Quine tries to convey his minimalist understanding of what it is to use one's sentences to make assertions, his task is therefore not one of theorizing about it, but of elucidating it. His strategy for elucidating it is to invite us, his readers, to start with our present understanding of our use of sentences to make assertions and then to *subtract* our doubtful or confused assumptions about what is required for such use. Subtracting these assumptions from our understanding of our use of sentences to make assertions, we get Quine's resolutely minimalist understanding of language use as the difference.

I. TRUTH AND DISQUOTATION

Passages in which Quine tries to elucidate his understanding of language use are scattered throughout his writings, but there is an unusually high concentration of them in *Word and Object*, §6. Here

---

* I presented a shorter version of this paper at the "*Word and Object* at 50" symposium at the 2010 Pacific Division meetings of the American Philosophical Association. The symposium was conceived, organized, and chaired by Janet Levin; the other symposium speakers were Michael Friedman, James Higginbotham, and Thomas Ricketts. Thanks to Janet Levin for inviting me to speak at the symposium, thereby prompting me to write this paper, and for her comments about it after the symposium. Thanks also to Tom Ricketts for suggesting that I discuss meaning holism in the paper; to Tom, Michael Friedman, and Jim Higginbotham for their comments at the symposium, and to Barry Stroud for a helpful reference. For reading and commenting on previous drafts of the paper, I thank Kate Abramson, Susan Blake, Matt Carlson, Adam Leite, Mark Kaplan, Mike Koss, Andrew McAninch, Blakely Phillips, Steve Wagner, and Joan Weiner.

[1] W. V. Quine, *Word and Object* (Cambridge: MIT, 1960)

0022-362X/11/0811/617–632

Quine emphasizes, "What reality is like is the business of scientists, in the broadest sense, painstakingly to surmise."[2] This is Quine's scientific naturalism: we use sentences to surmise, or theorize, but we have no grip on *what* to theorize apart from scientific method, loosely understood as what shapes our sense of what best explains the phenomena that interest us. In other words, scientific method is our last arbiter of truth.[3] The truth of a given sentence cannot in general be defined in terms of scientific method, however. First, it is unlikely that there is a unique best total theory, each of whose sentences is true.[4] Second, even if there *were* such a theory we could not use it to define truth for sentences of other theories, since "there is in general no sense in equating a sentence of a theory θ with a sentence S given apart from θ. Unless pretty firmly and directly conditioned to sensory stimulation, a sentence S is meaningless except relative to its own theory; meaningless intertheoretically."[5] Quine concludes that truth can only be defined disquotationally:

> Where it makes sense to apply 'true' is to a sentence couched in the terms of a given theory and seen from within the theory, complete with its posited reality. Here there is no occasion to invoke even so much as the imaginary codification of scientific method. To say that the statement 'Brutus killed Caesar' is true, or that 'The atomic weight of sodium is 23' is true, is in effect simply to say that Brutus killed Caesar, or that the atomic weight of sodium is 23.[6]

In this passage Quine invites us to acknowledge that even if meaning is indeterminate, we can use our sentences to make assertions, including, but of course not only, assertions about when a truth predicate applies to them.

Quine's talk of applying 'true' only to a sentence couched in the terms of a given theory and seen from within the theory may seem to suggest that he is committed to "a relativistic doctrine of truth—rating the statements of each theory as true for that theory, and brooking no higher criticism."[7] But he resolutely rejects any such relativism about truth:

> The saving consideration is that we continue to take seriously our own particular aggregate science, our own particular world-theory or loose total fabric of quasi-theories, whatever it may be. Unlike Descartes, we

---

[2] *Ibid.*, p. 22

[3] *Ibid.*, p. 23.

[4] *Ibid.*

[5] *Ibid.*, p. 24.

[6] *Ibid.*

[7] *Ibid.*

> own and use our beliefs of the moment, even in the midst of philosophizing, until by what is vaguely called scientific method we change them here and there for the better. Within our own total evolving doctrine, we can judge truth as earnestly and absolutely as can be; subject to correction, but that goes without saying.[8]

It is no accident that Quine places this ringing rejection of relativism at the very end of §6, in which he previews his reasons for rejecting the standard assumptions that synonymy is an objective relation and that the norms of epistemic justification are prior to and independent of our best current scientific theories. By subtracting these assumptions from our understanding of our use of sentences to make assertions, Quine thinks, we get a clearer, more useful understanding of our use of language as the difference.

II. ENTRENCHED CRITICISMS

According to Jaegwon Kim, however, Quine's subtractions leave us with nothing worthy of the label "epistemic justification." Kim argues that epistemic justification "manifestly is normative," since it concerns whether or not it is permissible, reasonable, or epistemically responsible for a person to hold a given belief.[9] The problem, according to Kim, is that Quine's causal-dispositional account of the relationship between "theory" (the totality of sentences we are disposed to affirm) and "evidence" (impacts at our nerve endings) is descriptive, not normative, and hence tells us nothing about when it is permissible, reasonable, or epistemically responsible for a person to hold a given belief.[10]

There is a grain of truth to Kim's criticism: if we focus exclusively on Quine's causal-dispositional account of the relationship between "theory" and "evidence," then normativity drops out of the picture. But if we focus on what Quine tells us about epistemology from the standpoint of our own evolving theory, we can eke out a residual, context-dependent sort of normativity. For instance, in the passage from *Word and Object*, §6, in which Quine introduces his disquotational account of truth, he writes:

> [The] statement[s] 'Brutus killed Caesar' is true, or that 'The atomic weight of sodium is 23' is true...are justifiable only by supplementing observation with scientific method...[11]

---

[8] *Ibid*., pp 24–25.

[9] Jaegwon Kim, "What Is 'Naturalized Epistemology'?" in *Philosophical Perspectives*, II (1988): 381–405, at p 383.

[10] *Ibid*., pp. 388–89

[11] Quine, *op cit*., p 24; Bredo Johnsen, "How to Read 'Epistemology Naturalized'," this JOURNAL, CII, 2 (February 2005). 78–93, at p 83

In the same spirit, Quine would agree with Gilbert Harman that an inference from one belief to another is epistemically justified if we accept it as part of our best current scientific theorizing.[12]

One might object that if this is all there is to say about justification, one would not be able to distinguish between *taking* oneself to be justified on a given occasion and *being* justified on that occasion, and hence one would not be able to see oneself as guided by norms of justification when one judges or asserts. This objection seems especially powerful if we focus on Quine's view that there is no sharp distinction between theoretical questions and pragmatic questions of language policy.[13] On this view, a scientist working from within the standpoint of his own current theory may weigh pragmatic questions of language policy in deciding what to believe. For questions of language policy, however, there can be no objective, inquirer-independent standards of right or wrong. It is therefore unclear what, if anything, remains of the apparently indispensible distinction between a speaker's taking herself to be justified on a given occasion and her being justified on that occasion. But if nothing of that distinction remains, then nothing remains of our idea of epistemic standards that *guide* us in our efforts to judge or assert in accord with the truth. How, then, on Quine's view, can there so much as *be* an act of judging or asserting?

A similar challenge arises from the meaning-theoretic side. H. P. Grice and P. F. Strawson argue that Quine's naturalism about meaning leads to the absurd conclusion that our sentences are meaningless. "If talk of sentence-synonymy is meaningless," they argue, "then it seems that talk of sentences having a meaning at all must be meaningless too."[14] This criticism, which is directed at Quine's arguments in "Two Dogmas of Empiricism," looks even more compelling after one reads chapter two of *Word and Object*, in which Quine argues in detail for his indeterminacy thesis. For if translation is objectively indeterminate, then the meanings of one's own sentences are objectively indeterminate. It apparently follows that when I try to "use" a sentence to assert, for instance, that a rabbit is not a rabbit part, I will inevitably fail, since my words have no determinate meaning.

This conclusion may seem inescapable if we read Quine's account of meaning as a holistic theory of the meanings of sentences. According to Michael Dummett, for example, "Quine...inclines to a holistic view of language, according to which the full explanation of the meaning of any sentence will involve an explanation of the entire language."[15] On this reading, "Meaning…becomes for Quine something essentially ineffable. We cannot say what meaning our language, or any part of it, has, since any attempt to do so would only yield a further constituent of that to which meaning is to be assigned."[16] If we agree that on Quine's view meanings are ineffable, it is unclear how we can accept his view and still take ourselves to use our sentences to make assertions.

---

[12] Gilbert Harman, "Inferential Justification," this JOURNAL, LXXIII, 17 (Oct. 7, 1976): 570–71.

[13] Quine, *op. cit.*, p. 271.

[14] H. P. Grice and P. F. Strawson, "In Defense of a Dogma," *The Philosophical Review*, LXV, 2 (April 1956): 141–58, at p. 146.

We get a similar result without relying on Quine's indeterminacy thesis or his supposed holism about sentence-meaning if we accept a claim that is central to Saul Kripke's presentation of Wittgenstein's remarks about meaning: our words are meaningful only if there are facts that determine how they ought to be used. As Kripke argues, it apparently follows from any dispositional theory of meaning, including Quine's, that our assertions of sentences are mere "jack-in-the-box" reactions to stimuli, and hence that nothing determines how our sentences ought to be used.[17] John McDowell agrees: if our utterance of a sentence is just a causal-dispositional reaction to stimuli, he reasons, then it is no more than a "mere brute meaningless sounding off."[18] In short, if our words are meaningful only if there are facts that determine how they ought to be used, then Quine's dispositional account of meaning apparently leads to the absurd conclusion that our words are meaningless, and so we cannot use them to make assertions.

III. THE FALLACY OF SUBTRACTION AND EXPLICATION

These entrenched criticisms of Quine's philosophy rest on a failure to grasp his resolutely minimalist conception of language use. To see why, recall first that Quine aims to subtract confused and mistaken assumptions about meaning and justification from our understanding of our use of sentences to make assertions. Grice and Strawson's main inference—"If talk of sentence-synonymy is meaningless, then it seems that talk of sentences having a meaning at all must be meaningless too"[19]—in effect reasserts some of the very views of meaning that Quine aims to subtract from our understanding of language use. Since Quine rejects Grice and Strawson's requirements on meaningfulness,

---

[15] Michael Dummett, *The Seas of Language* (New York: Oxford, 1993), p. 139.

[16] Dummett, *Frege: Philosophy of Language*, 2nd ed. (Cambridge. Harvard, 1981),

[17] Saul Kripke, *Wittgenstein on Rules and Private Language: An Element* (Cambridge: Harvard, 1982), p. 23.

[18] John McDowell, "Wittgenstein on Following a Rule," *Synthese*, 3 (March 1984): 325–63, at p 336

[19] Grice and Strawson, *op. cit.*, p. 146.

GENERAL LIBRA
UNIVERSITY

he sees no reason to accept their conclusion that our use of sentences to make assertions is incompatible with there being no objective synonymy relation between sentences. Their main inference is an instance of what Quine calls the "fallacy of subtraction," whereby it is argued that if we start with a sentence that we use to make assertions and we subtract our assumption that sentence-synonymy is an objective relation, we are left with a meaningless sentence, that is, one that we cannot use to make assertions.[20]

Quine sees no incompatibility between his thesis that translation is indeterminate and our use of sentences to make assertions. He therefore spends very little time on the fallacy of subtraction. In *Word and Object*, he even shows some exasperation at having to deal with the fallacy at all. He had already done so, he thought, in earlier publications, especially his influential 1948 paper "On What There Is," where he wrote, "I feel no reluctance toward refusing to admit meanings, for I do not thereby deny that words and statements are meaningful."[21]

The fallacy of subtraction will escape us if, like Dummett and many others, we suppose that Quine's account of meaning is a holistic theory of sentence-meaning. For then we will believe that Quine is committed to the conclusion that sentence-meaning is ineffable, and hence that we cannot use our sentences to make assertions. This reading of Quine is mistaken, however, as Hilary Putnam points out in "Meaning Holism";[22] like Wittgenstein's notorious suggestion that "meaning is use," Quine's account of meaning and his thesis of the indeterminacy of translation do not amount to a substantive theory of sentence-meaning, but to a denial of the appropriateness of theorizing about sentence-meaning. We have confirmation of Putnam's reading in Quine's approval of it,[23] as well as in Quine's charge that Grice and Strawson commit the fallacy of subtraction.

Kripke's jack-in-the-box objection also commits a version of the fallacy of subtraction. Since Quine rejects Kripke's requirements on using one's words to make assertions, he sees no reason to accept Kripke's conclusion that we can use our sentences to make assertions only if there is more to our linguistic behavior than jack-in-the-box reactions to stimuli.

---

[20] Quine, *op. cit.*, p. 206.

[21] Quine, "On What There Is," in *From a Logical Point of View*, rev. 2nd ed (Cambridge: Harvard, 1980), pp. 1–19, at p. 11

[22] Hilary Putnam, "Meaning Holism," in Lewis Edwin Hahn and Paul Arthur Schilpp, eds., *The Philosophy of W. V. Quine* (La Salle, IL: Open Court, 1986), pp. 406–26, see pp. 423–24, where Putnam quotes a passage from Warren Goldfarb, "I Want You to Bring Me a Slab: Remarks on the Opening Sections of the *Philosophical Investigations*," *Synthese*, LVI, 3 (September 1983): 265–82.

[23] Quine, "Reply to Putnam," in Hahn and Schilpp, eds., *op. cit.*, pp 427–31

One might believe that a proper analysis of our concept of meaning implies that Grice, Strawson, and Kripke's requirements for meaningfulness are correct, and hence that Quine cannot deny their requirements without replacing our concept of meaning with a different one, thereby simply ceasing to talk about meaning at all.

Quine's reply to this sort of objection is well known: we should not worry about preserving or capturing all features of our ordinary concept of meaning. We should focus instead on formulating clear, useful, and explanatory theories. What is less well known is that Quine explicitly endorses a method for this purpose. Following Carnap, he calls it the "method of explication." To *explicate* a linguistic expression *e* that one finds useful in some ways yet problematic in others is to decide to use, in place of *e*, a different linguistic expression *e′* that preserves and clarifies what one takes to be useful about *e* yet avoids what one takes to be the problems with *e*.[24] In Quine's view, the proper question to ask is not whether his rejection of Grice, Strawson, and Kripke's requirements for using sentences to make assertions is faithful to our concepts of meaning, language use, and assertion. The proper question is whether Quine's rejection of those supposed requirements furthers our goals as inquirers. If the answer to this latter question is "yes," then Quine's rejection of the requirements is vindicated, and the meaning-theoretic objections fall flat.

Many philosophers ignore or dismiss this reply, because they believe that Quine's method of explication presupposes his indeterminacy thesis and thereby begs the question against conceptual analysis. Following Carnap, however, Quine emphasizes that a synonymy constraint would be out of place for the sorts of paraphrases that matter to science, and to philosophy when it is viewed as continuous with science, "even if the notion of synonymy...were in the best of shape."[25] If we seek above all to provide the best explanations of observed phenomena, we should use our best current explications of terms previously in use and not concern ourselves with conceptual analyses of them, even if we suppose that synonymy is an objective relation. This methodological point stands on its own. It needs no support from Quine's indeterminacy thesis. In fact, the direction of support is the other way around: Quine's case for his indeterminacy thesis depends on his explication of meaning as what is determined by the totality of speech dispositions.

One might think that there is a better explication of meaning than Quine's, an explication that does not have the result that translation is indeterminate. Perhaps there is. But the possibility of such an explication is no objection to Quine's argument that if we explicate meaning as he proposes, synonymy is not an objective relation, and meaning is not normative. Nor is it an adequate response to Quine's charge that many philosophers commit the fallacy of subtraction. For Quine is surely within his rights to point out that if we adopt his explication of meaning, we can coherently reject Grice, Strawson, and Kripke's requirements for meaningfulness while still acknowledging that we use our sentences to make assertions.

---

[24] Quine, *Word and Object*, §53.

[25] *Ibid.*, p. 208, see also p 159.

IV. LANGUAGE USE AND SEARLE'S OBJECTION

To remind us what it is to use our words to make assertions, and to try to convince us that such language use is compatible with his indeterminacy thesis, Quine displays particular uses of sentences and invites us to acknowledge that they are uses of sentences to make assertions. Recall, for instance, that when Quine introduces his disquotational account of truth in §6 of *Word and Object*, he simply disquotes and uses his sample sentences, 'Brutus killed Caesar' and 'The atomic weight of sodium is 23'. His method here is the same as in his earlier paper "Notes on the Theory of Reference," in which he emphasizes that Tarski's method of defining truth:

> serves to endow...'true-in-*L*'...with every bit as much clarity, in any particular application, as is enjoyed by the particular expressions of *L* to which we apply [it]. Attribution of truth in particular to 'Snow is white', for example, is every bit as clear to us as attribution of whiteness to snow.[26]

Here and in other passages, Quine elucidates his conception of language use by putting language use on display in a way that reminds us of its peculiar clarity and transparency.

In "Ontological Relativity," Quine tries once again to elucidate his resolutely minimalist notion of language use. This time he displays language use as a response to a sophisticated new version of the Grice-Strawson challenge to the coherence of his indeterminacy thesis. He notes that the indeterminacy thesis applies to one's own language, and that this seems to imply that "there is no difference on any terms, interlinguistic or intralinguistic, objective or subjective, between referring to rabbits and referring to rabbit parts or stages."[27] In the objector's voice, he writes, "Surely this is absurd, for it would imply that there is no difference between the rabbit and each of its parts or stages."[28]

---

[26] Quine, "Notes on the Theory of Reference," in *From a Logical Point of View*, pp. 130–38, at p. 138.

[27] Quine, "Ontological Relativity," in *Ontological Relativity and Other Essays* (New York: Columbia, 1969), pp. 26–68, at p. 47.

Quine states the supposed absurdity in this way precisely in order to show that there is some mistake in the reasoning that leads to it. To recognize the absurdity of the conclusion that there is no difference between the rabbit and each of its parts or stages is also, and thereby, to recognize that we can and do deny that *there is no difference between the rabbit and each of its parts or stages*—in just those words.

John Searle argues that we cannot deny this, in just those words, without committing ourselves to rejecting Quine's indeterminacy thesis.[29] According to Searle, we must presuppose that the references of our own words are determinate when we use them to say such things as that rabbits are distinct from rabbit parts or rabbit stages. And Searle believes that Quine himself acknowledges the necessity of this presupposition when Quine notes the absurdity of the conclusion that there is no difference between the rabbit and each of its parts or stages. For at that critical point in his argument, Quine seems to acknowledge that we cannot simultaneously take ourselves to use our words to make assertions and reject the determinacy of reference. If Searle is right, Quine's resolutely minimalist account of language use is internally incoherent.

There is a striking structural similarity between Searle's objection and the other meaning-theoretic objections discussed above. In each case some requirement for using our words to make assertions is laid down; once the requirement is laid down, it is argued that Quine's account of meaning violates it and thereby implies that we cannot make assertions. This structural similarity should lead us to wonder whether Searle's argument also commits something like the fallacy of subtraction. But in Searle's case, what might the fallacy be?

---

[28] *Ibid.*, pp. 47–48

[29] John R. Searle, "Indeterminacy, Empiricism, and the First Person," this JOURNAL, LXXXIV, 3 (March 1987): 123–46; the argument I summarize in the text is presented on pp. 130–33. In a similar vein, though without claiming that Quine's position is incoherent, Barry Stroud notes that it is natural to think that "the sentence 'The word "rabbit" refers to rabbits' will explain the reference of that word to us only if we know what rabbits are. And someone's uttering that disquotational sentence would amount to saying that the word "rabbit" refers to rabbits only if he is referring to rabbits in using (not mentioning) that word in that sentence." Stroud, "Quine on Exile and Acquiescence," reprinted in *Meaning, Understanding, and Practice* (New York: Oxford, 2000), pp. 151–69, at p. 161. This natural thought poses a serious difficulty for Quine, since his indeterminacy thesis applies to one's own words, not just those of other speakers, and undermines what appears to be our presupposition that our word 'rabbit' determinately refers to rabbits. How then can we use our word 'rabbit' to say that it refers to rabbits? Many other philosophers endorse something like Searle's reasoning, which is now deeply entrenched in the literature. For instance, Ernie Lepore and Kirk Ludwig reject Donald Davidson's version of Quine's thesis that meaning and reference are indeterminate as incoherent on the grounds that an interpreter "cannot regard his own language as subject to the same indeterminacy that he is supposed to think of his subject's language as being subject to." Lepore and Ludwig, *Donald Davidson: Meaning, Truth, Language, and Reality* (New York: Oxford, 2005), p. 385.

It would not help to point out that in Quine's view, even though reference is indeterminate, every acceptable translation of my words yields some assignment or other of references to them. For if this remark is supposed to bear directly on our understanding of language use, it threatens to spoil our sense of the peculiar *transparency* of language use, including the use on display when we recognize the absurdity of saying that there is no difference between the rabbit and each of its parts or stages. To take this threat seriously is to endorse Searle's central assumption that we can use our words to make assertions only if reference is determinate. And if we endorse Searle's central assumption, we will see no way to resist Searle's conclusion that Quine's indeterminacy thesis is incoherent.

Despite its intuitive appeal, however, Searle's central assumption is mistaken. Quine explicates reference as satisfaction, defined in the first instance disquotationally. Thus explicated, reference is exclusively a relation between words and things that we define by using our words. Reference in this sense is not at issue when one uses one's nonsemantic vocabulary in the resolutely minimalist sense that is pragmatically indispensible to Quine's philosophy. Insofar as we are thus using, and not also mentioning, a given string of nonsemantic words, no particular relation between those words and things is asserted, presupposed, or implied.

There is textual evidence that this is exactly how Quine would respond to the sort of problem that Searle raises. Quine apparently thinks we can block the absurd implication that there is no difference between the rabbit and each of its parts or stages by displaying our language use without talking of reference at all. He attempts this in the following passage:

> Toward resolving this quandary, begin by picturing us at home in our language, with all its predicates and auxiliary devices. This vocabulary includes "rabbit," "rabbit part," "rabbit stage,"...; also the two-place predicates of identity and difference, and other logical particles. In these terms we can say in so many words that this is...a rabbit and that a rabbit part, this and that the same rabbit, and this and that different parts. *In just those words.*[30]

Quine's aim here is to remind us that we use our words to say, for instance, that this is a rabbit and that a rabbit part, this and that the same rabbit, and this and that different parts. By reminding us of indexical applications of our sentences, Quine also reminds us that our use of words is responsive to publicly observable conditions, and thereby hints at a final, crucial point: his minimalist account of language use in no way conflicts with his view that meaning is public and indeterminate.

---

[30] Quine, "Ontological Relativity," p. 48.

V. EPISTEMOLOGY AGAIN

Let us revisit the objection that what Quine tells us about epistemology from the standpoint of our own evolving theory conflates a speaker's taking a given assertion to be justified with its being justified. How might Quine respond to this objection? Let us grant, for the reasons just explained, that Quine can assume that we are each always in a position to express conditions under which our sentences are true by simply using them. This is not to equate their truth with our current beliefs about whether they are true. After I have made a particular judgment, I may later look back on that judgment and regard it as mistaken. From the standpoint of my current best theory, I can criticize another theory, including one of my own past best theories. At the time of adopting a theory, of course, I cannot distinguish from within my subjective point of view between having justification and taking myself to have justification. But when I reflect on one of my previous judgments, I may see something that I overlooked at the time, something that from my current point of view seems to me to undermine my previous confidence that my judgment was justified. Perhaps I come to think, for example, that while I took myself to have been exercising the best kind of scientific judgment on a given occasion, in fact I was not doing so. One need not believe that there are universal, context-independent standards for exercising scientific judgment in order to accept the possibility of this kind of retrospective assessment of whether a judgment is justified.

One might think that this reply does not yield an objective purchase on the notion of justification. For the best we can do, according to the Quinean position just stated, is to judge justification from our best current theoretical point of view. To evaluate a past judgment, we would first have to choose a translation *T* from our past idiolect to our present one and then evaluate past judgments, as mapped by *T* into our current idiolect, by our current standards. At any given time, we have no way to distinguish between our current inclination to take a judgment as justified, on the one hand, and its being justified, on the other.

To regard this objection as decisive one must suppose that there is some privileged standpoint from which to evaluate particular claims to have justification—a standpoint on justification that is independent of any particular theory that we adopt. In Quine's view, however, there is no such standpoint. Scientific method, unsupported by ulterior controls, is "the last arbiter of truth,"[31] and we have no grip on scientific method apart from our own evolving scientific theories. Hence, for Quine it is only from the standpoint of our best current theories that we can judge truth "as earnestly and absolutely as can be."[32]

This aspect of Quine's philosophy—the uncompromising core of his scientific naturalism—can seem puzzling. For if there is no legitimate standpoint on justification that is independent of any particular scientific theories we accept, then there are no substantive general principles for evaluating or justifying our acceptance of particular scientific statements or the theories of which they are a part. How then can an inquirer's acceptance of particular statements and theories be epistemically reasonable?[33]

Quine's response to this question combines an application of his method of explication with an application of his method of elucidation by subtraction.[34] To get on with our inquiries, he observes, we need a practical grasp of when it is appropriate to adopt a theory and what it is to justify, defend, and revise our assertions from the standpoint of a given theory. The best way to attain such a grasp is to immerse oneself, by study and practice, in the relevant scientific disciplines. Such immersion yields a practical and theoretical grasp of the disciplines that in no way depends on one's reflection about or acceptance of any traditional philosophical theory of justification. Quine concludes that an inquirer who has immersed herself in some scientific disciplines and wishes above all to get on with her inquiries has no need for a fully general, discipline-independent account of what it is for her acceptance of a given theory or statement in any of her disciplines to be justified.

One might challenge this conclusion on the grounds that an inquirer who starts by fully endorsing some scientific disciplines she has learned might subsequently be led by her own skeptical reflections to conclude that by the standards implicit in those scientific disciplines, and contrary to what she used to assume, she actually has no justification for any of her scientific assertions. Such an inquirer might arrive at the conclusion that to understand how she can justify any assertion she needs to have a fully general understanding of how human knowledge is possible at all—an understanding that she despairs of ever attaining.[35] That an inquirer could arrive at a skeptical conclusion in this way may appear to discredit Quine's conclusion that an inquirer has no need for a fully general, discipline-independent account of what it is for her acceptance of a given theory or statement in any of her disciplines to be justified.

---

[31] Quine, *Word and Object*, p. 23.

[32] *Ibid.*, p. 25.

[33] This question is similar to one raised by Gideon Rosen, who suggests that Quine has no good answer to what Rosen calls the "Authority Problem for Naturalized Epistemology"—the problem of giving "some sort of principle for telling the authoritative practices from the rest." See Rosen's review of Penelope Maddy, *Naturalism in Mathematics*, in *British Journal for the Philosophy of Science*, L, 3 (September 1999): 467–77, at p. 471

[34] Quine also develops a theory of how our sentences are associated with sensory stimulation by the mechanism of conditioned response. But such descriptions do not convey what it is to use one's sentences to accept theories or to justify, defend, or revise statements from the standpoint of a theory one has accepted. When Quine tries to convey his minimalist understanding of what it is to use one's sentences in these familiar ways, his task is not to theorize about it, but to elucidate it.

This challenge rests on the assumption that there are standards—in particular, certain requirements for epistemological justification—that traditional philosophers see in our scientific practices but that Quine's minimalist account of those practices fails to capture. As I noted in section III above, Quine counters other such challenges to his philosophy by judicious applications of his method of explication. In epistemology, or what little is left of it when its questions are posed from the standpoint of an inquirer who is engaged in scientific inquiry, Quine's scientific naturalism commits him to explicating and elucidating epistemic terms in a resolutely minimalist way that preserves applications of the epistemic terms that inquirers find clear and useful without leaving inquirers vulnerable to traditional skeptical arguments. Like inquiry itself, this task may never be completed. New methods in science may create new vulnerabilities to old skeptical arguments that call for new explications. Once we replace the method of conceptual analysis with the method of explication, however, there is no general reason to worry that our inquiries can be decisively undermined from within by traditional skeptical doubts.[36] Hence, if our primary goal is to get on with our inquiries, as Quine assumes, we are free to explicate the phrase "epistemically reasonable" in such a way that our judgments and inferences can be epistemically reasonable even if there are no substantive general principles for evaluating them.[37]

---

[35] According to Stroud, if we take the traditional skeptic's reasoning seriously we will come to realize that "We want an account of our knowledge of the world that would make all of it intelligible to us all at once," but that such an account is impossible, since "we cannot consider all our knowledge of the world all at once and still see it as knowledge." Stroud, "Skepticism and the Possibility of Knowledge," this JOURNAL, LXXXI, 10 (October 1984): 545–51, at p. 551.

[36] Dirk Koppelberg offers a roughly similar response on Quine's behalf to traditional skepticism, though without mentioning Quine's methods of explication and elucidation by subtraction. Koppelberg writes, "What a [Quinean] naturalist denies is the skeptic's presupposition of the intelligibility of a completely external perspective on our knowledge....if we resist the temptation to adopt an external standpoint, in favor of trying to explain our knowledge from within our evolving and changing theory of the world, the significance of philosophical skepticism might yet decline. When it amounts to the insight that our traditional notions of knowledge and objectivity stand in need of

One might object that this explication abandons our concept of what it is to be epistemically reasonable and replaces it with a radically different one that has nothing to do with epistemology. But this objection can appear to us to discredit Quine's epistemology only if we cling dogmatically to the claim that if we abandon the traditional epistemologist's assumption that there are substantive general principles for evaluating and justifying assertions, we lose our grip on what it is to justify, defend, and revise our assertions from the standpoint of a given scientific discipline. To challenge this claim, all Quine can do is invite us to subtract the assumption that there are substantive general principles for evaluating and justifying assertions from the practical and theoretical grasp of what it is to justify, defend, and revise assertions in a given scientific discipline that we achieve by study and practice of that discipline. Subtracting this assumption from our practical and theoretical grasp of what it is to justify theories or statements in the scientific disciplines we have learned or developed, we get Quine's resolutely minimalist understanding of justification as the difference.

The resulting account of justification will not satisfy an epistemologist who wishes above all to capture and preserve her traditional epistemological preconceptions, including her assumption that there are substantive general principles for justifying assertions. But the objections of such an epistemologist are no more challenging to Quine's efforts to elucidate our practical and theoretical grasp of what it is to justify theories or statements in the scientific disciplines we have learned or developed than are Grice, Strawson, Kripke, or Searle's objections to Quine's efforts to elucidate our use of sentences to make assertions. The two sorts of objections are, indeed, at root identical.[38]

---

revision, naturalism can be seen to offer a promising way with this task at hand." Koppelberg, "Why and How to Naturalize Epistemology," in Robert B. Barrett and Roger F. Gibson, eds., *Perspectives on Quine* (Cambridge: Blackwell, 1990), pp. 200–11, quotation from p. 210 In his "Comment on Koppelberg," in *ibid.*, p. 212, Quine expresses his "thanks and admiration" for Koppelberg's analysis of his epistemological position.

[37] Consider in this light the last sentence of "Two Dogmas of Empiricism": "Each man is given a scientific heritage plus a continuing barrage of sensory stimulation; and the considerations which guide him in warping his scientific heritage to fit his continuing sensory promptings are, where rational, pragmatic." Quine, *From a Logical Point of View*, pp. 20–46, at p. 46.

[38] Recall Quine's claim that the dogma that there is an analytic-synthetic distinction and the dogma of reductionism "are, indeed, at root identical" *Ibid.*, p. 41 The two

Quine's response to them is to reject their assumptions and to rely on our pragmatically prior grasp of what it is to assert and justify theories and sentences from the standpoint of a scientific discipline we endorse.

VI. THE LAST WORD

For just this reason, Thomas Nagel sees Quine's philosophy as a form of subjectivism that "shrink[s] from the apparently audacious pretensions of human thought and tend[s] to collapse its content into its grounds."[39] Nagel observes that we treat the results of our best current first-order theorizing as the last word—as absolutely objective and unconditionally authoritative for us right now—unless and until we see reason to revise the results, and thereby to arrive at a new result that we treat as the latest last word. Quine can easily paraphrase this preliminary observation into the resolutely minimalist language of his scientific naturalism. But Nagel goes further. He sees Quine's scientific naturalism (wrongly, I think, for reasons I have given above) as an attempt to underwrite the authority of our judgments with a psychological theory of our needs and linguistic dispositions. Against this, Nagel stresses (rightly, I think), "The authority of the most fundamental kinds of thought reveals itself only from inside each of them and cannot be underwritten by a theory of the thinker."[40] Surprisingly, he goes on to theorize about what it is for our thoughts to be authoritative "from the inside." He claims, for instance, that we do not understand the authority of some of our thoughts unless we can give substantive content to the truism that "To reason is to think systematically in ways anyone looking over my shoulder ought to be able to recognize as correct."[41] And he argues that this truism has substantive content only if certain rationalist assumptions are true. More generally, in his view, we can reason systematically, debate our reasons with others, and use our language to make assertions that count for us now as the last word on a given topic only if our sentences have objective contents or meanings and our judgments are justified by universally binding reasons. Nagel concludes that since Quine tries to do without these rationalist assumptions, he fails to explain the phenomenon of the last word.

My reply to Nagel's criticisms of Quine is similar to my replies to the criticisms that I discussed above. The apparent appeal of Nagel's rationalistic requirements on using one's sentences to express one's

---

dogmas that Quine identifies are of a piece with the meaning-theoretic and epistemological objections to Quine's philosophy that I have considered in this paper

[39] Thomas Nagel, *The Last Word* (New York: Oxford, 1997), p. 7.

[40] *Ibid.*, p. 26

[41] *Ibid.*, p. 5

last word on a given topic is an illusion sustained by a flawed philosophical methodology, one that places conceptual analysis at center stage, and hence fails to keep truth in focus. The best way to avoid such illusions is to replace the method of conceptual analysis with the method of explication, and to keep one's philosophy focused on truth, not on concepts or justification. To reason and to judge, "we can never do better than occupy the standpoint of some theory or other, the best we can muster at the time."[42] Hence, it is fruitless to look for universally binding justifications for our reasoning and judging, beyond those reasons we already offer when we apply scientific method to arrive at our best current theories. We must therefore reject Nagel's rationalistic requirements on reasoning and judging, along with all philosophical dreams of a perspective on reasoning and judging that is somehow higher or firmer than our own current theories. Only when our rejections of such misguided philosophical theorizing are complete can we attain a clear view of what it is to take a bit of discourse as the last word—to judge truth "as earnestly and absolutely as can be."[43] In short, it is Quine, not Nagel, who gets the last word.

GARY EBBS

Indiana University, Bloomington

---

[42] Quine, *Word and Object*, p. 22.

[43] *Ibid.*, p. 25.

COMMENTS AND CRITICISM

WILLIAMS ON SUPERVALUATIONISM AND LOGICAL REVISIONISM[*]

Central to discussion of supervaluationist accounts of vagueness is the extent to which they require revisions of classical logic and to the extent that they do, whether those revisions are objectionable. In an important recent *Journal of Philosophy* article, J. R. G. Williams presents a powerful challenge to the orthodox view that supervaluationism is objectionably revisionary.[1] Williams argues both that supervaluationism is nonrevisionary and that even if it were, those revisions would be unobjectionable. This note shows that his arguments for both claims fail.

I. THE CASE FOR REVISIONISM

Williams begins with a model-theoretic characterization of consequence in a supervaluationist setting. A supervaluationist model M is a quadruple $<D_M, \Delta_M, R_M, I_M>$. D_M is the domain, Δ_M is a set of delineations, R_M is an accessibility relation on D_M, and I_M is an interpretation function from expressions and delineations to classical extensions. A notion of truth relative to models and delineations—truth at M,d—is defined in the standard way. This much is exactly analogous to a possible-worlds semantics (without an actual world) for a modal language.

Supervaluationist semantics departs from standard modal semantics in its characterizations of truth and consequence. Say that a sentence Φ is *supertrue* in M iff Φ is true at M,d, for every delineation $d \in \Delta_M$. Truth in a model M is then identified with supertruth in M. Since some sentences are neither supertrue nor superfalse in some models, bivalence fails. Both local and global consequence relations are definable within this framework:

> $\Gamma \vDash_{\text{local}} \Phi$ iff, for every model M and delineation $d \in \Delta_M$, if every member of Γ is true at M,d then Φ is true at M,d.

---

[*] I would like to thank Will Bynoe, Mike Caie, Dorothy Edgington, Simon Hewitt, and Ian Rumfitt for discussion. Thanks also to Robbie Williams for written comments on an earlier version. This material was presented at the King's Metaphysics Society and the 2010 London-Berkeley graduate conference; thanks to the audiences at both events for their comments. This research was made possible by funding from the Arts and Humanities Research Council and a Royal Institute of Philosophy Jacobsen Fellowship; I am very grateful to both organizations.

[1] J. R. G. Williams, "Supervaluationism and Logical Revisionism," this JOURNAL, CV, 4 (April 2008): 192–212.

0022-362X/11/0811/633–641

> $\Gamma \models_{\text{global}} \Phi$ iff, for every model M, if every member of Γ is supertrue in M, then Φ is supertrue in M.

Given the identification of truth with supertruth, there are well-known reasons to identify consequence proper with global consequence, rather than its local counterpart.[2] To these, Williams adds a powerful new argument.[3] So let us set aside local consequence: so far as supervaluationism is concerned, global consequence is consequence.

The purported counterexamples to classical logic arise with the introduction of a 'Definitely' operator D, akin to a necessity operator $\Box$ in possible-worlds semantics:

> $\mathsf{D}\Phi$ is true at M,d iff Φ is true at M,d, for every delineation $d \in \Delta_M$ R_M-accessible from d.

The following results all hold, providing apparent counterexamples to the respective classical rules of proof:

Contraposition:
- $p \models_{\text{global}} \mathsf{D}p$
- $\neg\mathsf{D}p \not\models_{\text{global}} \neg p$

Conditional proof:
- $p \models_{\text{global}} \mathsf{D}p$
- $\not\models_{\text{global}} p \supset \mathsf{D}p$

Argument by cases:
- $\models_{\text{global}} p \vee \neg p$
- $p \models_{\text{global}} \mathsf{D}p$
- $p \models_{\text{global}} \neg\mathsf{D}p$
- $\not\models_{\text{global}} \mathsf{D}p \vee \neg\mathsf{D}p$

Reductio:
- $p \wedge \neg\mathsf{D}p \models_{\text{global}} \bot$
- $\not\models_{\text{global}} \neg(p \wedge \neg\mathsf{D}p)$

Since none of these results holds for local consequence, logical revisionism could be avoided by identifying that with consequence proper. We have seen that this is not Williams's approach. His strategy is to argue that the present supervaluationist framework is inadequate; the results all fail in a more satisfactory setting.

II. WILLIAMS'S CASE AGAINST REVISIONISM

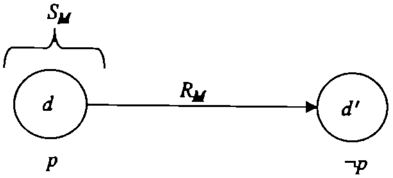
Williams begins by observing that plausible semantic analyses of linguistic phenomena other than vagueness, specifically comparatives, have been proposed that require delineations other than those relevant to the determination of (super)truth-value.[4] A fully general semantic theory for a vague language may well have to incorporate such delineations. To accommodate this, let an *extended* model M be just like a supervaluationist model but with a fifth element: $<D_M, \Delta_M, R_M, I_M, S_M>$, where $S_M \subseteq \Delta_M$. The elements of S_M are the *sharpenings* of M. Supertruth in M is then redefined as truth at M,d, for all sharpenings $d \in S_M$. The sharpenings are the privileged subset of delineations that contribute to determining truth-value. Extended models are just like standard possible-worlds models but with a set of actual worlds. If the quantifiers over models in our characterization of global consequence range over all extended models, then all the results in the previous section fail.

---

[2] The following argument for the equation of consequence with global consequence is a variant on one due to Timothy Williamson, *Vagueness* (New York: Routledge, 1994), p. 148. According to the supervaluationist, truth on an interpretation is supertruth in a supervaluationist model. Since consequence is preservation of truth on all interpretations, it follows that consequence is global consequence. For dissent, see Achille C. Varzi, "Supervaluationism and Its Logics," *Mind*, CXVI, 463 (July 2007): 633–76; and Pablo Cobreros, "Supervaluationism and Logical Consequence: A Third Way," *Studia Logica*, XC, 3 (December 2008): 291–312. Since Williams identifies supervaluationist consequence with global consequence, we can safely bracket this issue.

[3] Williams, *op. cit.*, §3.

To see this, note first that central to the purported counterexamples to the classical rules are the results $p \vDash_{\text{global}} \mathsf{D}p$ and $p \wedge \neg\mathsf{D}p \vDash_{\text{global}} \perp$. Neither holds in Williams's extended setting. For let M be an extended model with just two delineations d, d', whose only sharpening is d, where d R_M-accesses d', and such that p is true at M,d but not true at M,d':



Since d is the only sharpening and p is true at M,d: p is supertrue in M. Since d' is R_M-accessible from d, and p is not true at M,d': $\mathsf{D}p$ is not true at M,d. So: $\neg\mathsf{D}p$ is true at M,d and hence supertrue in M. So M is a countermodel to both $p \vDash_{\text{global}} \mathsf{D}p$ and $p \wedge \neg\mathsf{D}p \vDash_{\text{global}} \perp$.

The game is not yet up. Let the *admissible* extended models be those over which the quantifiers in our characterization of global consequence range. Let the *restricted-access* (RA) models be those where the sharpenings access only other sharpenings. It is crucial to Williams's case that the admissible models not be restricted to RA models. For suppose that p is supertrue in an RA model M. Then no sharpening accesses any delineation where p is not true. So $\mathsf{D}p$ is supertrue in M. Since M was arbitrary: $p \vDash_{\text{global}} \mathsf{D}p$ and, hence, $p \wedge \neg\mathsf{D}p \vDash_{\text{global}} \perp$ hold unless the admissible models include some non-RA models. A two-step strategy for reinstating the counterexamples now emerges. First step: argue that only RA models respect the intended sense of D. Second step: argue that admissible models must respect the intended sense of D.

---

[4] *Ibid.*, §2.

Williams pre-empts this strategy. Against the first step, he constructs a toy non-RA model which, he claims, does respect the intended sense of D.[5] He continues:

> Perhaps some inventive defender of the orthodox position could make a case that the toy model constructed above, and all others like it, are unfaithful to the intended sense of 'Definitely' (I have no idea what such a case would look like, but cannot rule it out).[6]

Section III provides just such a case. Against the second step, Williams argues that restricting the admissible models to RA models involves treating D as a logical constant.[7] Since the proper characterization of the logical constants is highly contentious, Williams concludes that the case for logical revisionism is weak at best. Section IV argues that the case for logical revisionism can sidestep this issue: the proper characterization of the logical constants is irrelevant. Section V closes by showing that even if Williams is right on both counts, a slightly different argument for logical revisionism is available that avoids his complaints. Furthermore, this last argument undermines Williams's case for regarding the purported revisions as unobjectionable.

III. REINSTATING REVISIONISM: FIRST STEP

This section argues that supervaluationists should regard only RA models as faithful to the intended sense of D. The argument rests on two explanatory demands on a satisfactory account of vagueness. The first is an account of borderline status and definiteness: what is it for a sentence to be definitely true, or an object to be definitely *F*?[8] The second is an account of borderline ignorance: if it is borderline whether *p*, then it seems somehow misguided to try and discover whether *p*. Why does borderline status interact with knowledge in this way? And if it does not, then why does it appear to?

Supervaluationism offers an account of borderline status, and hence definiteness, in terms of truth-value gaps: it is borderline whether *p* iff *p* is neither true nor false (in virtue of being true at some but not all sharpenings). Thus definiteness is analyzed in terms of more familiar semantic concepts. Note also that without this analysis of borderline cases in terms of truth-value gaps, the identification of truth with supertruth looks ill motivated. For that identification is motivated by the desire to avoid bivalence; unless borderline cases fall down a truth-value gap, it is mysterious why a nonbivalent semantics would be desirable. This analysis of borderline status naturally extends to an explanation of borderline ignorance: since knowledge implies truth, if p is borderline (neither true nor false), then it cannot be known whether p.[9] Thus the importance that D provide "an object-language reflection of the supervaluationist's notion of truth—'supertruth'."[10]

---

[5] *Ibid.*, §5.

[6] *Ibid.*, p. 205.

[7] *Ibid.*, §6.

[8] Some deny that any analysis of borderline status is possible. One example is David Barnett, "Is Vagueness *Sui Generis*?" *Australasian Journal of Philosophy*, LXXXVII, 1 (2009): 5–34. I take it that the lack of explanatory work for the concepts of definiteness and of borderline status beyond the study of vagueness and indeterminacy counts strongly against these views.

Neither of these explanations is available if non-RA models respect the intended sense of D. For in some non-RA models, p is both true and borderline (as the model diagrammed in the previous section shows); hence borderline status cannot be analyzed in terms of truth-value gaps, and borderline ignorance cannot be explained in terms of the untruth of borderline statements, if such models are faithful to the intended sense of D. If the intended sense of 'Definitely' permits these explanations, and if the identification of truth with supertruth is to be well motivated, then only RA models respect that sense.[11]

One might reply that $p \models_{\text{global}} \mathrm{D}p$ should fail anyway in the presence of higher-order vagueness. For an instance is $\mathrm{D}p \models_{\text{global}} \mathrm{DD}p$, which rules out borderline cases of definite cases. The objection is mistaken.

---

[9] In *Saving Truth from Paradox* (New York: Oxford, 2008), pp. 154–55, Hartry Field criticizes this account of borderline ignorance. But without an alternative, the supervaluationist should endorse it. I see only one alternative explanation for the appearance of borderline ignorance: although borderline status does not prevent knowledge, it does prevent definite knowledge, and one seems ignorant when one does not definitely know. Although this fits Williams's extended semantics that permits true borderline statements, it also strengthens the demand for an explanation of definiteness. Supposing I know that p, why does it matter whether I also definitely know that p? Why should an inability to definitely know make it (or make it appear) futile to try and discover whether p? Some account is surely owed, and it is hard to see where it might come from, if not from the analysis of definiteness. I see two options. (1) It is just a primitive fact about vagueness that borderline status makes it appear futile to try to know. (2) Definite knowledge is the goal of assertion, and borderline status creates the appearance of ignorance by making it in-principle illegitimate to assert that p when p is known to be borderline. Neither approach is satisfactory without an account of definiteness: why should definite knowledge, rather than mere knowledge, be the goal of assertion? Simply postulating primitive and inexplicable relationships between definiteness and other concepts is an unsatisfying approach to discharging the supervaluationist's explanatory obligations.

[10] Williams, *op. cit.*, p. 192.

[11] It does not strictly follow that only RA models respect the intended sense of D. The argument provides only the following necessary condition for doing so: no sharpening accesses any delineation where any supertruth is untrue. The restriction of admissible models to those that meet this constraint will reinstate the counterexamples to the classical rules. So we can safely ignore this complication.

If $p \vDash_{\text{global}} \mathsf{D}p$, then $\mathsf{D}p$, $\mathsf{DD}p$, $\mathsf{DDD}p, \ldots$ are all true in any model where p is true. But this is consistent with the falsity of $\mathsf{D}p$, $\mathsf{DD}p$, $\mathsf{DDD}p, \ldots$ and untruth of the S4 principle $\mathsf{D}p \supset \mathsf{DD}p$ in models where p is borderline (and hence untrue). By identifying truth with supertruth with clear truth, supervaluationism collapses the definite, definitely definite, definitely definitely definite, and so on, cases into the cases. But it does not make higher-order borderline cases logically impossible, as the objection assumes.

I conclude that supervaluationists should regard only RA models as faithful to the intended sense of D if they are to offer their customary semantic analysis of definiteness and explanation of borderline ignorance, as well as the point of identifying truth with supertruth. Since no alternative explanation is forthcoming, supervaluationists should regard only RA models as faithful to the intended sense of D.

IV. REINSTATING REVISIONISM: SECOND STEP

Should the admissible models be restricted to those faithful to the intended sense of D? Expressions whose interpretation is held constant across admissible models when characterizing logical consequence are called *logical constants*. So our question becomes: is D a logical constant? Williams claims that not only is the proper characterization of the logical constants too controversial to provide a firm basis on which to rest the case for revisionism, but even the proper application of competing accounts to particular cases is disputed. This is surely correct. So let us set aside the question of whether D is a logical constant and consider instead the following two questions: (1) Why are the interpretations of some expressions held fixed across admissible models? (2) What is an investigation of global consequence supposed to provide? We address each in turn.

Why hold fixed the interpretation of $\wedge$ when investigating the logical behavior of conjunction? Why not permit models that interpret $\wedge$ as disjunction? The answer is that these deviant interpretations are irrelevant to our interests. We want to know what follows from the conjunction of p and q, not from the truth of the string of symbols '$p \wedge q$'. Regardless of whether $\wedge$ is a logical constant, or whether logical consequence requires constancy in its interpretation across admissible models, if we want to know how conjunction contributes to the correctness of inferences, then we should investigate what consequence relation results when $\wedge$ is interpreted as conjunction in all admissible models. The point is that deductive inference involves the manipulation of contents, not strings of symbols; it involves discovering what must be the case, given other beliefs and suppositions as to what is the case. Some consequence relation provides the standard of correctness for this activity, regardless of whether that relation is logical consequence proper or whether the inferences in question are strictly logical. So if we want to know how to reason with conjunctive contents, then we should, on pain of simply changing the topic, investigate consequence relations that result from interpreting $\wedge$ as conjunction in all admissible models.

Let us turn to our second question: what should our investigation of global consequence provide? In part, we want an account of correct reasoning in a vague language. But we also want to know how to reason about vagueness: what follows from the supposition that p is borderline? The sole purpose of the D operator is to allow us to make such suppositions, to allow expression of definitized contents in a language with supervaluationist semantics. Thus our investigation of global consequence should provide an account of the contribution of D to the correctness of inferences. Classical predicate calculus and model theory provide a powerful tool for studying reasoning with predicative, conjunctive, disjunctive, negated, and quantified contents in a precise language. Supervaluationist model theory is supposed to provide a similar tool for studying reasoning with those same kinds of contents in a vague language. With the addition of D, we acquire the further capacity to study reasoning with contents concerning definiteness.

Combining these answers to our two questions yields the following: if we want to know how we should reason about definiteness, then we should investigate consequence relations where the admissible models are faithful to the intended sense of D. It is simply irrelevant to this investigation whether D is a logical constant, or whether any such consequence relation is logical consequence. We do want to know how to reason about definiteness in a vague language. Given the argument of the previous section, we should therefore investigate consequence relations that result from restricting the admissible models to the RA models. But then the initial results all hold; each provides a counterexample to the claim that all classically correct inferential patterns are legitimate in a language with supervaluationist semantics and the resources for expressing definiteness. Whether or not this is a revision to logic, the supervaluationist should not reason classically about definiteness.

V. A BETTER CASE FOR REVISIONISM

Suppose that my arguments in the preceding sections all fail. Then because not all admissible models are RA models: $p \nvDash_{\text{global}} \mathsf{D}p$ and $p \wedge \neg\mathsf{D}p \nvDash_{\text{global}} \bot$. We can show that supervaluationism is revisionary nonetheless. The reason is that nothing prevents the introduction of a truth operator via the following clause:

$\mathsf{T}p$ is true at M,d iff p is true (that is, supertrue) in M.

The initial results all hold under a uniform substitution of D for T: T induces new counterexamples to the classical rules. Since the semantic clause for T is given in terms of (super)truth in a model directly, there is no scope to avoid this by tinkering with the underlying space of sharpenings or an accessibility relation on them. Setting disputes about D to one side, the equivalence of $\mathsf{T}p$ to the truth of p at all and only the sharpenings is mandated by the identification of truth with supertruth and adoption of a global consequence relation. The new counterexamples to the classical rules thus follow directly from the key supervaluationist claim to retain LEM without bivalence. Furthermore, it is in terms of truth, rather than definiteness, that much discussion of the revisionary implications of supervaluationist semantics has been conducted.[12,13]

One might object that Liar-like paradoxes provide independent reason to doubt that the classical rules hold in full generality anyway, for languages that contain their own truth predicate. But since T is a sentential operator and not a predicate, we cannot use it to construct a Liar sentence or similar. The objection therefore fails.

I close by noting that T undermines Williams's case against regarding deviations from the classical rules as revisionary to classical inferential

---

[12] See, for example, Vann McGee and Brian McLaughlin, "Review of Williamson," *Linguistics and Philosophy*, XXI, 2 (April 1998): 221–35; and "Logical Commitment and Semantic Indeterminacy. A Reply to Williamson," *Linguistics and Philosophy*, XXVII, 1 (February 2004): 122–36; as well as Williamson, "Reply to McGee and McLaughlin," *Linguistics and Philosophy*, XXVII, 1 (February 2004): 113–22.

[13] In an appendix, Williams offers a proof that global and classical consequence coincide. It is invalid for languages containing T. To show this, first we need an account of classical consequence. A standard possible-worlds model M for a modal language is just one of our extended models with a single sharpening $\bullet$. Call any such model a *classical* model. Then we have: $\Gamma \vDash_{\text{classical}} \Phi$ iff, for every classical model M, if every member of Γ is true at $M,\bullet$ then Φ is true at $M,\bullet$. Equivalently: classical consequence is the global consequence relation that results from restricting the admissible models to the classical models. Williams now has to show that if $\Gamma \nvDash_{\text{global}} \Phi$, then $\Gamma \nvDash_{\text{classical}} \Phi$. He proceeds as follows:

> Suppose $\Gamma \nvDash_{\text{global}} \Phi$. Then for some extended model M: every member of Γ is supertrue in M and Φ is not supertrue in M. So for some sharpening $s \in S_M$: every member of Γ is true at M,s and Φ is not true at M,s. Let M^* be the extended model that differs from M only by substituting $\{s\}$ for S_M. M^* is a classical model where $\bullet = s$. By construction, every member of Γ is true at $M^*,\bullet$, but Φ is not true at $M^*,\bullet$. So M^* is a countermodel to: $\Gamma \vDash_{\text{classical}} \Phi$. Discharging our initial supposition: if $\Gamma \nvDash_{\text{global}} \Phi$, then $\Gamma \nvDash_{\text{classical}} \Phi$.

This proof fails in a language containing T. The problem is that although Γ are all true at M,s and Φ is not true at M,s, it does not follow that both (i) Γ are all true at M^*,s, and (ii) Φ is not true at M^*,s. To see this, let $\Gamma = \{\neg\mathsf{T}p\}$ and let $\Phi = \neg p$. Although there are extended models in which both $\neg\mathsf{T}p$ is supertrue and $\neg p$ is not supertrue, there is no classical model where this is so. For if $\neg p$ is not true at $M^c,\bullet$ for some classical model M^c, then p is true at $M^c,\bullet$; so $\mathsf{T}p$ is true at $M^c,\bullet$; hence $\neg\mathsf{T}p$ is not true at $M^c,\bullet$. Williams's proof fails because. $\neg\mathsf{T}p \nvDash_{\text{global}} \neg p$, but $\neg\mathsf{T}p \vDash_{\text{classical}} \neg p$

practice, as opposed to classical logical theory. Let $\vDash_{\text{global}}$ be the global consequence relation obtained by allowing admissible non-RA models; let $\vDash^{+}_{\text{global}}$ be the global consequence relation obtained by restricting admissible models to RA models. Williams argues that $\vDash^{+}_{\text{global}}$ does not involve deviation from classical inferential practice by appeal to:

CP* If $\Gamma, \Phi \vDash_{\text{global}} \Psi$, then: $\Gamma \vDash^{+}_{\text{global}} \Phi \supset \Psi$.

His idea is that, in order to show that supervaluationism mandates revisions to inferential practice, "we would have to show that inferential practice mandates moving from $\vDash^{+}_{\text{global}}$-valid but $\vDash_{\text{global}}$-invalid arguments, to the corresponding conditional. No such case has been made."[14] But in a language containing T, even CP* fails:

$p \vDash_{\text{global}} \mathsf{T}p$, but: $\nvDash^{+}_{\text{global}} p \supset \mathsf{T}p$.

Even if ordinary inferential practice involves no more than drawing conditional conclusions from $\vDash_{\text{global}}$-valid arguments, supervaluationism is revisionary of that practice if, as argued above, $\vDash^{+}_{\text{global}}$ is the standard for deductive correctness when reasoning using D. Contra Williams, supervaluationism does lead to logical revisionism.

NICHOLAS K. JONES

Merton College, University of Oxford

---

[14] Williams, *op cit*, p. 210.

BOOK REVIEWS

Kurt Gödel: Essays for His Centennial. SOLOMON FEFERMAN, CHARLES PARSONS, and STEVEN G. SIMPSON, eds. New York: Cambridge University Press, 2010. x + 374 p. Cloth $90.00.

Kurt Gödel was born on April 28, 1906. The centennial of his birth led to several scholarly meetings and publications to commemorate and celebrate the event. A meeting organized by the Association for Symbolic Logic (ASL) at the Université du Québec à Montréal (May 17–21, 2006) featured three plenary lectures on Gödel by Jeremy Avigad, Sy-David Friedman, and Akihiro Kanamori, as well as a special session on Gödel's philosophy of mathematics with lectures by Steve Awodey, John Burgess, and William Tait. But renewed attention to Gödel's work was in full swing even before then, as a consequence of the masterful completion (achieved slowly but surely) of the five volumes of Gödel's *Collected Works* under the general editorship of Solomon Feferman. Indeed, the 2004 ASL annual meeting at Carnegie Mellon University featured a special session organized by the co-editors of the *Collected Works.* The papers from this latter conference were published in a special issue of *The Bulletin of Symbolic Logic* in 2005.

The volume under review contains all the papers presented at the ASL meeting in 2006 in Montréal together with those that were included in the special issue of *The Bulletin of Symbolic Logic.* Three additional articles reprinted from other sources complete the collection. This yields a total of seventeen papers of outstanding quality divided into four major sections: (a) General; (b) Proof Theory; (c) Set Theory; (d) Philosophy of Mathematics. The authors are all well-known experts in mathematical logic and/or history and philosophy of logic and/or mathematics who share—it goes without saying—a deep knowledge of Gödel's work, its background, and its ramifications. They include Mark van Atten, Jeremy Avigad, Steve Awodey, John Burgess, Andrew W. Carus, Martin Davis, John W. Dawson, Cheryl A. Dawson, Solomon Feferman, Warren Goldfarb, Sy-David Friedman, Akihiro Kanamori, Juliette Kennedy, Peter Koellner, Charles Parsons, Wilfried Sieg, Stephen G. Simpson, and William Tait.

In a short Preface, the editors briefly comment on the rationale for the collection:

> [We] aimed at a collection that would cover Gödel's work in the fields represented by the Association for Symbolic Logic, thus mathematical

0022-362X/11/0811/642–646

> logic, especially proof theory and set theory, and philosophy of mathematics. For this reason other aspects of Gödel's work were omitted, in particular his contribution to general relativity, his philosophical reflection on relativity theory, and other parts of his philosophical thought. We have not aimed at comprehensiveness even in the fields within our purview. Nonetheless we hope that this volume is a fitting tribute to a great logician and will be of assistance to those studying his work and developing his ideas in the future. (ix–x)

The first section of the book contains an informative article by Feferman chronicling the complex stages that led from its inception in the late 1970s to the completion of the *Collected Works* in 2004. The article by John Dawson and Cheryl Dawson focuses on the work that went into classifying Gödel's *Nachlass* and gives a very good idea of the materials contained in it, as well as how much still needs to be done in transcribing many of his notebooks.

The second section is on proof theory and opens with Jeremy Avigad's article, "Gödel and the Mathematical Tradition." In this paper Avigad offers a reading of an interesting and productive tension in Gödel's intellectual output. Avigad notices that Gödel wrote polemically against finitist and formalist philosophies of mathematics, and yet almost the totality of his logical work can be seen as a contribution to those programs. Surprisingly, Avigad mentions the proof of completeness for first-order logic as the only exception to this characterization and does not include Gödel's work on the consistency of the Axiom of Choice (AC) and the Generalized Continuum Hypothesis (GCH) with the axioms of Zermelo-Fraenkel (ZF) or Bernays-Gödel (BG) set theory, but his rationale is that the latter are finitist relative-consistency results. The following article, authored by Sieg, gives a contextual analysis of the Gödel-Herbrand correspondence showing, among other things, that the Herbrand-Gödel notion of computable function is in reality due to Gödel. We then have two articles by Tait, one on Gödel's reading of Gentzen's first consistency proof for arithmetic and the other on Gödel's notion of intuition. The former is an extremely informative and detailed article on how Gödel's reading of Gentzen's first consistency proof (in his "Lecture at Zilsel's" (1938)) reformulates the proof in terms of what is now known as the no-counterexample interpretation. The latter article is a probing investigation of Gödel's notion of finitism and intuition and argues that Gödel's notion of finitism is not unstable, but rather underwent a smooth evolution. In addition, Tait argues that Kant's notion of arithmetic intuition can be captured by Primitive Recursive Arithmetic (PRA) and that it differs from Gödel's notion of finitism. Simpson's article ("The Gödel Hierarchy and Reverse Mathematics") argues that far from putting an end to the axiomatic study of theories, Gödel's incompleteness theorems are at the source of logical research that shows a "large amount of logical regularity and structure" (110). One such program is "reverse mathematics," which addresses the question, "What axioms are needed to prove particular mathematical results?" Simpson gives a very effective presentation of its main aims and achievements. Burgess ("On the Outside Looking In: A Caution about Conservativeness") discusses the philosophical implications of Gödel's incompleteness theorems and ties together two areas that are often discussed separately: the relevance of these theorems for Hilbert's program and the debates on whether minds are machines. The connection, it is argued, is that a line of objection to the argument that minds cannot be machines (an argument put forward by Lucas and Penrose, among others) displays a structural similarity to the situation according to which, while provability in primitive recursive arithmetic and "finitistic" provability might indeed coincide, this can only be seen "from the outside"; in particular, the finitist cannot have access to this fact "from the inside." These reflections are then applied to insightfully evaluate several claims made on behalf of partial realizations of Hilbert's program (including many from the reverse mathematics literature).

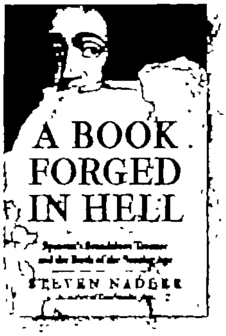
The section on set theory contains three articles by Akihiro Kanamori, Sy-David Friedman, and Peter Koellner. Kanamori's lengthy article presents an account of Gödel's engagement with set theory that brings to bear both historical and mathematical considerations to describe the immense distance Gödel had to cover, conceptually and technically, in his reorientation of set theory. This is one of the best accounts of Gödel's work in set theory that I have ever come across. Friedman describes the results of recent research in set theory which constitute generalizations of Gödel's universe of constructible sets, with special focus on the problem of combining the mathematical power of V=L (namely, the claim that the universe of sets V coincides with the class of sets L obtainable by means of certain "definable" operations) with the consistency power of large cardinals. Koellner's rich article starts from the question of "absolute undecidability" posed by Gödel. A theory like Peano Arithmetic (PA) has undecidable statements; for instance, the statement expressing the consistency of PA cannot be proved or disproved in PA. However, the consistency of PA can be derived from other evident principles through evident rules of inference. Gödel asked whether there are statements that are absolutely undecidable—specifically, statements that are undecidable relative to any justified set of axioms—and favored different answers at different times. Koellner first describes Gödel's development of the notion of absolute undecidability; he then reviews an impressive amount of recent work in set theory relevant to the issue and argues that "there is not currently a good argument to the effect that a given sentence is absolutely undecidable" (190).

The last section of the book is most likely to interest the greater number of readers of this JOURNAL. It opens with Martin Davis's historical account of Gödel's developing views on Hilbert's program and Platonism ("What Did Gödel Believe and When Did He Believe It?"). Davis sees in Gödel a progressive disenchantment with Hilbert's program and a definite contrast between Gödel's dismissal of set-theoretic Platonism in 1933 and his later Platonist views. Warren Goldfarb, in "On Gödel's Way In: The Influence of Rudolf Carnap," traces how Gödel's early work on completeness and incompleteness was articulated within the context of Carnap's *Investigations on General Axiomatics* and how, in turn, Gödel's work influenced Carnap's work in the early 1930s on logical syntax. The relationship between Gödel and Carnap is also the subject of Awodey and Carus's paper ("Gödel and Carnap"). In addition to a very clear analysis of the shortcomings of the Carnapian project in *Investigations on General Axiomatics*, they discuss and evaluate Gödel's objections to Carnap's project of mathematics as syntax of language, and they argue that Gödel's objections are not compelling. In "On the Philosophical Development of Kurt Gödel," Mark van Atten and Juliette Kennedy engage in a thorough historical and philosophical excavation of the role Husserl played in Gödel's thinking and what is distinctive of Gödel's phenomenological inspiration. They show that Husserl's influence can be detected in a number of published works by Gödel and that Husserl provided Gödel with an account of subjectivity (transcendental idealism) that promised to bridge his deeply held commitments to rationality and Platonism. This paper is to be recommended as the best treatment of the relation between Gödel and Husserl but also for the various insights it yields concerning Gödel's attitude to Leibniz, Kant, and German idealism. It usefully complements work by Parsons and Tieszen on the same subject. Parsons's article, "Platonism and Mathematical Intuition in Kurt Gödel's Thought," is the oldest anthologized in this book (it was published in 1995). Parsons argues that Gödel's theory of intuition is posterior and not intrinsically connected to his Platonism; in the process, he provides a fine account of the nature of Gödel's notion of intuition and highlights the problems that it was meant to solve. The last article in the collection, by Donald A. Martin, is titled "Gödel's Conceptual Realism." The special type of Platonism held by Gödel in set theory emphasizes realism about the concept of set, as opposed to realism about sets themselves (while also defending the latter). Martin carefully probes what Gödel means by "concept" and "perception of the concept of set," thereby guiding the reader through the complexities of Gödel's realism in set theory. He further proposes an interpretation of Gödel's concept realism by reading "concept of set" as "concept of set structures."

In a review of Volumes IV and V of Gödel's *Collected Works*, I asked, "What more do we know that we could not have known without the correspondence?" (*Notre Dame Journal of Formal Logic*, XLV, 2 (2004): 109–25, at p. 112). I claimed there that our understanding of Gödel's work and context had been extended substantially by the wealth of new material that Gödel's *Nachlass* had yielded and that the *Collected Works* had put at our disposal. The present collection reflects the deep impact of the *Collected Works* on Gödel scholarship. I speak of Gödel scholarship in a broad sense meant to include detailed studies of the historical context of Gödel's work, the philosophical and exegetical analyses of Gödel's standpoint on a variety of issues (realism, finitism, intuition, and so on), as well as the formulation of contemporary programs in mathematical logic that can be seen as probing Gödel's claims and conjectures by either generalizing, confirming, or disconfirming them (interaction between notions of absolute undecidability and recent work in set theory, generalizations of Gödel's notion of constructible sets, work in reverse mathematics, and so on). It is of course regrettable that in the short space of this review, I was not able to detail, let alone engage, the arguments, both exegetical and technical, raised in the articles in the collection. On the other hand, it would have been invidious to select a small group of the articles for discussion. Nonetheless, I hope to have conveyed that this is a first-rate collection that belongs on the bookshelf of anyone interested in the history and philosophy of logic and mathematics, and mathematical logic. Preferably—and I will certainly do so—the volume should be placed next to Gödel's *Collected Works*, to which it provides a wonderful complement.

PAOLO MANCOSU

University of California, Berkeley



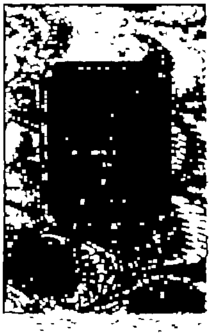
A Book Forged in Hell

Spinoza's Scandalous Treatise and the Birth of the Secular Age

Steven Nadler

"*A Book Forged in Hell* is more than just an excellent and highly readable introduction to one of the most important texts of philosophy and political thought. Steven Nadler provides an eloquent portrait of Spinoza's treatise, placing it firmly in its historical, religious, political, and philosophical setting."
—Jonathan Israel, author of *A Revolution of the Mind*

Cloth $29.95 978-0-691-13989-0



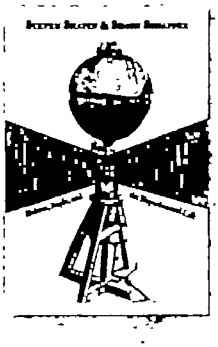
Locke on Personal Identity

Consciousness and Concernment

Galen Strawson

"Galen Strawson proposes an original and provocative interpretation of Locke's treatment of personal identity. Strawson makes his case with characteristic depth, insight, ingenuity, and clarity. This engagingly written work should be of great interest to historians of modern philosophy and to all philosophers working on personal identity."
—Don Garrett, New York University

Cloth $35.00 978-0-691-14757-4



With a new introduction by the authors

Leviathan and the Air-Pump

Hobbes, Boyle, and the Experimental Life

Steven Shapin and Simon Schaffer

"*Leviathan and the Air-Pump* [is] the most influential text in our field since Thomas Kuhn's *Structure of Scientific Revolutions*."
—James Secord, *Isis*

"This is simply one of the most original, enjoyable, and important books published in the history of science in recent years."
—Owen Hannaway, *Technology and Culture*

Paper $29.95 978-0-691-15020-8

CAMBRIDGE

New and Noteworthy Titles in Philosophy from Cambridge University Press!

Should Race Matter?
Unusual Answers
to the Usual Questions
David Boonin
$99.00 Hb 978-0-521-76086-7
$34.99 Pb 978-0-521-14980-8
422 pp.

The Evolution of Modern Metaphysics
Making Sense of Things
A. W. Moore
The Evolution of Modern Philosophy
$110.00 Hb 978-0-521-85111-4
$39.99 Pb 978-0-521-61655-3
690 pp

Peace Education
How We Come to Love and Hate War
Nel Noddings
$90.00 Hb 978-0-521-19382-5
$24.99 Pb 978-1-107-65872-1
190 pp

The Cambridge Companion to Nozick's *Anarchy, State, and Utopia*
Ralf M. Bader *and*
John Meadowcroft
Cambridge Companions to Philosophy
$90.00 Hb 978-0-521-19776-2
$29.99 Pb 978-0-521-12002-9
332 pp

Now in Paperback!
The Cambridge History of Eighteenth-Century Philosophy, 2 Volume Pb Boxed Set
Knud Haakonssen
$90.00 2 Pb bks.
978-0-521-28088-4 1,424 pp

Cambridge Applied Ethics

Ethics and War
An Introduction
Steven P. Lee
$95.00 Hb 978-0-521-89883-6
$29.99 Pb 978-0-521-72757-0
342 pp

Ethics and the Media
An Introduction
Stephen J. A. Ward
$90.00 Hb 978-0-521-88964-3
$28.99 Pb 978-0-521-71816-5
298 pp

Insurrection and Intervention
The Two Faces of Sovereignty
Ned Dobos
$85.00 Hb 978-0-521-76113-0
244 pp

Nietzsche's *On the Genealogy of Morality*
A Critical Guide
Simon May
Cambridge Critical Guides
$99.00 Hb 978-0-521-51880-2
356 pp

Kant and Cosmopolitanism
The Philosophical Ideal of World Citizenship
Pauline Kleingeld
$90.00 Hb 978-0-521-76418-6
232 pp

Kant, Religion, and Politics
James DiCenso
$95.00 Hb 978-1-107-00933-2
302 pp

Spinoza's Geometry of Power
Valtteri Viljanen
$90.00 Hb 978-1-107-00780-2
208 pp

Agro-Technology
A Philosophical Introduction
R. Paul Thompson
Cambridge Introductions to Philosophy and Biology
$90.00 Hb 978-0-521-11797-5
$28.99 Pb 978-0-521-13375-3
258 pp

Consciousness and the Self
New Essays
JeeLoo Liu *and*
John Perry
$90.00 Hb 978-1-107-00075-9
270 pp

Kuhn's Evolutionary Social Epistemology
K. Brad Wray
$95.00 Hb 978-1-107-01223-3
244 pp

Logical Dynamics of Information and Interaction
Johan van Benthem
$90.00 Hb 978-0-521-76579-4
386 pp

Proof Analysis
A Contribution to
Hilbert's Last Problem
Sara Negri *and*
Jan von Plato
$90.00 Hb 978-1-107-00895-3
278 pp

Prices subject to change

www.cambridge.org/us/philosophy

CAMBRIDGE
UNIVERSITY PRESS